Faster Algorithms for Testing under Conditional Sampling

Moein Falahatgar
mfalahat@ucsd.edu

Ashkan Jafarpour
ashkan@ucsd.edu

Alon Orlitsky
alon@ucsd.edu

Venkatadheeraj Pichapathi
dheerajpv7@gmail.com

Ananda Theertha Suresh
asuresh@ucsd.edu

University of California, San Diego

April 17, 2015

Abstract

There has been considerable recent interest in distribution-tests whose run-time and sample requirements are sublinear in the domain-size k. We study two of the most important tests under the conditional-sampling model where each query specifies a subset S of the domain, and the response is a sample drawn from S according to the underlying distribution.

For identity testing, which asks whether the underlying distribution equals a specific given distribution or ϵ-differs from it, we reduce the known time and sample complexities from $\widetilde{\mathcal{O}}(\epsilon^{-4})$ to $\widetilde{\mathcal{O}}(\epsilon^{-2})$, thereby matching the information theoretic lower bound. For closeness testing, which asks whether two distributions underlying observed data sets are equal or different, we reduce existing complexity from $\widetilde{\mathcal{O}}(\epsilon^{-4}\log^5 k)$ to an even sub-logarithmic $\widetilde{\mathcal{O}}(\epsilon^{-5}\log\log k)$ thus providing a better bound to an open problem in Bertinoro Workshop on Sublinear Algorithms [Fisher, 2014].

Keywords: Property testing, conditional sampling, sublinear algorithms

1 Introduction

1.1 Background

The question of whether two probability distributions are the same or substantially different arises in many important applications. We consider two variations of this problem: *identity testing* where one distribution is known while the other is revealed only via its samples, and *closeness testing* where both distributions are revealed only via their samples.

As its name suggests, identity testing arises when an identity needs to be verified. For example, testing whether a given person generated an observed fingerprint, if a specific author wrote an unattributed document, or if a certain disease caused the symptoms experienced by a patient. In all these cases we may have sufficient information to accurately infer the true identity's underlying distribution, and ask whether this distribution also generated newly-observed samples. For example, multiple original high-quality fingerprints can be used to infer the fingerprint structure, and then be applied to decide whether it generated newly-observed fingerprints.

Closeness testing arises when we try to discern whether the same entity generated two different data sets. For example, if two fingerprints were generated by the same individual, two documents were written by the same author, or two patients suffer from the same disease. In these cases, we do not know the distribution underlying each data set, but would still like to determine whether they were generated by the same distribution or by two different ones.

Both problems have been studied extensively. In the hypothesis-testing framework, researchers studied the asymptotic test error as the number of samples tends to infinity, [see Ziv, 1988, Unnikrishnan, 2012, and references therein]. We will follow a more recent, non-asymptotic approach. Two distributions p and q are ϵ-*far* if

$$||p - q||_1 \geq \epsilon.$$

An *identity test* for a given distribution p considers independent samples from an unknown distribution q and declares either $q = p$ or they are ϵ-far. The test's *error probability* is the highest probability that it errs, maximized over $q = p$ and every q that is ϵ-far from p. Note if p and q are neither same nor ϵ-far, namely if $0 < ||q - p||_1 < \epsilon$, neither answer constitutes an error.

Let $N_{\mathrm{id}}(k, \epsilon, \delta)$ be the smallest number of samples to identity test every k-element distribution with error probability $\leq \delta$. It can be shown that the sample complexity depends on δ mildly, $N_{\mathrm{id}}(k, \epsilon, \delta) \leq \mathcal{O}(N_{\mathrm{id}}(k, \epsilon, 0.1)) \cdot \log \frac{1}{\delta}$. Hence we focus on $N_{\mathrm{id}}(k, \epsilon, 0.1)$, denoting it by $N_{\mathrm{id}}(k, \epsilon)$.

This formulation was introduced by Goldreich and Ron [2000] who, motivated by testing graph expansion, considered identity testing of uniform distributions. Paninski [2008] showed that the sample complexity of identity testing for the uniform distributions is $\Theta(\epsilon^{-2}\sqrt{k})$. General identity testing was studied by Batu et al. [2001] who showed that $N_{\mathrm{id}}(k, \epsilon) \leq \widetilde{\mathcal{O}}(\epsilon^{-2}\sqrt{k})$, and recently Valiant and Valiant [2013] proved a matching lower bound, implying that $N_{\mathrm{id}}(k, \epsilon) = \Theta(\epsilon^{-2}\sqrt{k})$, where $\widetilde{\mathcal{O}}$ and later $\widetilde{\Theta}$ and $\widetilde{\Omega}$, hide multiplicative logarithmic factors.

Similarly, a *closeness test* takes independent samples from p and q and declares them either to be the same or ϵ-far. The test's *error probability* is the highest probability that it errs, maximized over $q = p$ and every p and q that are ϵ-far. Let $N_{\mathrm{cl}}(k, \epsilon, \delta)$ be the smallest number of samples that suffice to closeness test every two k-element distributions with error probability $\leq \delta$. Here too it suffices to consider $N_{\mathrm{cl}}(k, \epsilon) \stackrel{\mathrm{def}}{=} N_{\mathrm{cl}}(k, \epsilon, 0.1)$.

Closeness testing was first studied by Batu et al. [2000] who showed that $N_{\mathrm{cl}}(k, \epsilon) \leq \widetilde{\mathcal{O}}(\epsilon^{-4}k^{2/3})$. Recently Valiant [2011], Chan et al. [2014b] showed that $N_{\mathrm{cl}}(k, \epsilon) = \Theta(\max(\epsilon^{-4/3}k^{2/3}, \epsilon^{-2}\sqrt{k}))$.

1.2 Alternative models

The problem's elegance, intrinsic interest, and potential applications have led several researchers to consider scenarios where fewer samples may suffice. Monotone, log-concave, and m-modal distributions were considered in Rubinfeld and Servedio [2009], Daskalakis et al. [2013], Diakonikolas et al. [2015], Chan et al. [2014a], and their sample complexity was shown to decline from a polynomial in k to a polynomial in $\log k$. For example, identity testing of monotone distributions over k elements requires $\mathcal{O}(\epsilon^{-5/2}\sqrt{\log k})$ samples, and identity testing log-concave distributions over k elements requires $\widetilde{\mathcal{O}}(\epsilon^{-9/4})$ samples, independent of the support size k.

A competitive framework that analyzes the optimality for every pair of distributions was considered in Acharya et al. [2012], Valiant and Valiant [2013]. Other related scenarios include classification [Acharya et al., 2012], outlier detection [Acharya et al., 2014b], testing collections of distributions [Levi et al., 2013], testing for the class of monotone distributions [Batu et al., 2004], testing for

the class of Poisson Binomial distributions [Acharya and Daskalakis, 2015], testing under different distance measures [Guha et al., 2009, Waggoner, 2015].

Another direction lowered the sample complexity of all distributions by considering more powerful queries. Perhaps the most natural is the *conditional-sampling* model introduced independently in Chakraborty et al. [2013] and Canonne et al. [2014], where instead of obtaining samples from the entire support set, each query specifies a *query set* $S \subseteq [k]$ and the samples are then selected from S in proportion to their original probability, namely element i is selected with probability

$$P_S(i) = \begin{cases} \frac{p(i)}{p(S)} & i \in S, \\ 0 & \text{otherwise}, \end{cases}$$

where $p(S)$ is the probability of set S under p. Conditional sampling is a natural extension of sampling, and Chakraborty et al. [2013] describes several scenarios where it may arise. Note that unlike other works in distribution testing, conditional sampling algorithms can be adaptive, *i.e.,* each query set can depend on previous queries and observed samples. It is similar in spirit to the machine learning's popular *active testing* paradigm, where additional information is interactively requested for specific domain elements. Balcan et al. [2012] showed that various problems such as testing unions of intervals, testing linear separators benefit significantly from the active testing model.

Let $N^*_{\mathrm{id}}(k,\epsilon)$ and $N^*_{\mathrm{cl}}(k,\epsilon)$ be the number of samples required for identity- and closeness-testing under conditional sampling model. For identity testing, Canonne et al. [2014] showed that conditional sampling eliminates the dependence on k,

$$\Omega(\epsilon^{-2}) \le N^*_{\mathrm{id}}(k,\epsilon) \le \widetilde{\mathcal{O}}(\epsilon^{-4}).$$

For closeness testing, the same paper showed that

$$N^*_{\mathrm{cl}}(k,\epsilon) \le \widetilde{\mathcal{O}}(\epsilon^{-4}\log^5 k).$$

Chakraborty et al. [2013] showed that $N^*_{\mathrm{id}}(k,\epsilon) \le \mathrm{poly}(\log^* k, \epsilon^{-1})$ and designed a $\mathrm{poly}(\log k, \epsilon^{-1})$ algorithm for testing any label-invariant property. They also derived a $\Omega(\sqrt{\log\log k})$ lower bound for testing any label-invariant property.

An open problem posed by Fisher [2014] asked the sample complexity of closeness testing under conditional sampling which was partly answered by Acharya et al. [2014a], who showed

$$N^*_{\mathrm{cl}}(k,1/4) \ge \Omega(\sqrt{\log\log k}).$$

1.3 New results

Our first result resolves the sample complexity of identity testing with conditional sampling. For identity testing we show that

$$N^*_{\mathrm{id}}(k,\epsilon) \le \widetilde{\mathcal{O}}(\epsilon^{-2}).$$

Along with the information-theoretic lower bound above, this yields

$$N^*_{\mathrm{id}}(k,\epsilon) = \widetilde{\Theta}(\epsilon^{-2}).$$

For closeness testing, we address the open problem of Fisher [2014] by reducing the upper bound from $\log^5 k$ to $\log\log k$. We show that

$$N^*_{\mathrm{cl}}(k,\epsilon) \le \widetilde{\mathcal{O}}\big(\epsilon^{-5}\log\log k\big).$$

This very mild, double-logarithmic dependence on the alphabet size may be the first sub-poly-logarithmic growth rate of any non-constant-complexity property and together with the lower bound in Acharya et al. [2014a] shows that the dependence on k is indeed a poly-double-logarithmic.

Rest of the paper is organized as follows. We first study identity testing in Section 2. In Section 3 we propose an algorithm for closeness testing. All the proofs are given in Appendix.

2 Identity testing

In the following, p is a distribution over $[k] \stackrel{\text{def}}{=} \{1, \ldots, k\}$, $p(i)$ is the probability of $i \in [k]$, $|S|$ is the cardinality of $S \subseteq [k]$, p_S is the conditional distribution of p when S is queried, and n is the number of samples. For an element i, $n(i)$ is used to denote the number of occurrences of i.

This section is organized as follows. We first motivate our identity test using restricted uniformity testing, a special case of identity testing. We then highlight two important aspects of our identity test: finding a *distinguishing element* i and finding a *distinguishing set* S. We then provide a simple algorithm for finding a distinguishing element. As we show, finding distinguishing sets are easy for testing *near-uniform* distributions and we give an algorithm for testing near-uniform distributions. We later use the near-uniform case as a subroutine for testing any general distribution.

2.1 Example: restricted uniformity testing

Consider the class of distributions $\mathcal{Q}$, where each $q \in \mathcal{Q}$ has $k/2$ elements with probability $(1+\epsilon)/k$, and $k/2$ elements with probability $(1-\epsilon)/k$. Let p be the uniform distribution, namely $p(i) = 1/k$ for all $1 \le i \le k$. Hence for every $q \in \mathcal{Q}$, $||p - q||_1 = \epsilon$.

We now motivate our test via a simpler *restricted uniformity testing*, a special case of identity testing where one determines if a distribution is p or if it belongs to the class $\mathcal{Q}$.

If we know two elements i, j such that $q(i) = \frac{1+\epsilon}{k} > \frac{1}{k} = p(i)$ and $q(j) = \frac{1-\epsilon}{k} < \frac{1}{k} = p(j)$, it suffices to consider the set $S = \{i, j\}$. For this set

$$p_S(i) = \frac{p(i)}{p(i) + p(j)} = p_S(j) = \frac{p(j)}{p(i) + p(j)} = \frac{1/k}{2/k} = \frac{1}{2},$$

while

$$q_S(i) = \frac{q(i)}{q(i) + q(j)} = \frac{(1+\epsilon)/k}{(1+\epsilon)/k + (1-\epsilon)/k} = \frac{1+\epsilon}{2},$$

and similarly $q_S(j) = (1-\epsilon)/2$. Thus differentiating between p_S and q_S is same as differentiating between $B(1/2)$ and $B((1+\epsilon)/2)$ for which a simple application of the Chernoff bound shows that $\mathcal{O}(\epsilon^{-2})$ samples suffice. Thus the sample complexity is $\mathcal{O}(\epsilon^{-2})$ if we knew such a set S.

Next consider the same class of distributions $\mathcal{Q}$, but without the knowledge of elements i and j. We can pick two elements uniformly at random from all possible $\binom{k}{2}$ pairs. With probability $\ge 1/2$, the two elements will have different probabilities as above, and again we could determine whether root the distribution is uniform. Our success probability is half the success probability when S is known, but it can be increased by repeating the experiment several times and declaring the distribution to be non-uniform if one of the choices of i and j indicates non-uniformity.

While the above example illustrates tests for uniform distribution, for non-uniform distributions finding elements i, j can be difficult. Instead of finding pairs of elements, we find a distinguishing element i and a distinguishing set S such that $q(i) < p(i) \approx p(S) < q(S)$, thus when conditional

samples from $S \cup \{i\}$ are observed, the number of times i appears would differ significantly, and one can use Chernoff-type arguments to differentiate between `same` and `diff`. While previous authors have used similar methods, our main contribution is to design a information theoretically near-optimal identity test.

Before we proceed to identity testing, we quantify the Chernoff-type arguments formally using TEST-EQUAL. It takes samples from two unknown binary distributions p, q (without loss of generality assume over $\{0,1\}$), error probability δ, and a parameter ϵ and it tests if $p = q$ or $\frac{(p-q)^2}{(p+q)(2-p-q)} \geq \epsilon$. We use the chi-squared distance $\frac{(p-q)^2}{(p+q)(2-p-q)}$ as the measure of distance instead of ℓ_1 since it captures the dependence on sample complexity more accurately. For example, consider two scenarios: $p, q = B(1/2), B(1/2+\epsilon/2)$ or $p, q = B(0), B(\epsilon/2)$. In both cases $||p-q||_1 = \epsilon$, but the number of samples required to distinguish p and q in the first case is $\mathcal{O}(\epsilon^{-2})$, while in the second case $\mathcal{O}(\epsilon^{-1})$ suffice. However, chi-squared distance correctly captures the sample complexity as in the first case it is $\mathcal{O}(\epsilon^2)$ and in the second case it is $\mathcal{O}(\epsilon)$. While several other simple hypothesis tests exist, the algorithm below has near-optimal sample complexity in terms of ϵ, δ.

Algorithm TEST-EQUAL
Input: chi-squared bound ϵ, error δ, distributions $B(p)$ and $B(q)$.
Parameters: $n = \mathcal{O}(1/\epsilon)$.
Repeat $18 \log \frac{1}{\delta}$ times and output the majority:

1. Let $n' = \text{poi}(n)$ and $n'' = \text{poi}(n)$ be two independent Poisson variables with mean n.
2. Draw samples $x_1, x_2 \ldots x_{n'}$ from the first distribution and $y_1, y_2 \ldots y_{n''}$ from the second one.
3. Let $n_1 = \sum_{i=1}^{n'} x_i$ and $n_2 = \sum_{i=1}^{n''} y_i$.
4. If $\frac{(n_1-n_2)^2 - n_1 - n_2}{n_1+n_2-1} + \frac{(n_1-n_2)^2-n_1-n_2}{n'+n''-n_1-n_2-1} \leq \frac{n\epsilon}{2}$ then output `same`, else `diff`.

Lemma 1 (Appendix B.1)**.** *If $p = q$, then* TEST-EQUAL *outputs* `same` *with probability $1-\delta$. If $\frac{(p-q)^2}{(p+q)(2-p-q)} \geq \epsilon$, it outputs* `diff` *with probability $\geq 1-\delta$. Furthermore the algorithm uses $\mathcal{O}\left(\frac{1}{\epsilon} \cdot \log \frac{1}{\delta}\right)$ samples.*

2.2 Finding a distinguishing element i

We now give an algorithm to find an element i such that $p(i) > q(i)$. In the above mentioned example, we could find such an element with probability $\geq 1/2$, by randomly selecting i out of all elements. However, for some distributions, this probability is much lower. For example consider the following distributions p and q. $p(1) = \epsilon/2$, $p(2) = 0$, $p(i) = \frac{1-\epsilon/2}{k-2}$ for $i \geq 2$, and $q(1) = 0$, $q(2) = \epsilon/2$, $q(i) = \frac{1-\epsilon/2}{k-2}$ for $i \geq 2$. Again note that $||p-q||_1 = \epsilon$. If we pick i at random, the chance that $p(i) > q(i)$ is $1/k$, very small for our purpose. A better way of selecting i would be sampling according to p itself. For example, the probability of finding an element i such that $p(i) > q(i)$ when sampled from p is $\epsilon/2 \gg 1/k$.

We quantify the above idea next by using the following simple algorithm that picks elements such that $p(i) > q(i)$. We first need the following definition. Without loss of generality assume that the elements are ordered such that $p(1) \geq p(2) \geq p(3) \ldots \geq p(k)$.

Definition 2. *For a distribution p, element i is α-heavy, if $\sum_{i':i' \geq i} p(i') \geq \alpha$.*

As we show in proofs, symbols that are heavy (α large) can be used as distinguishing symbols easily and hence our goal is to choose symbols such that $p(i) > q(i)$ and i is α-heavy for a large value of α. To this end, first consider an auxiliary result that shows if for some non-negative values a_i, $\sum_i p(i)a_i > 0$, then the following sampling algorithm will pick an element x_i such that x_i is α_i-heavy and $a_{x_i} \geq \beta_i$. While several other algorithms have similar properties, the following algorithm achieves a good trade-between α and β (one of the tuples satisfy $\alpha\beta = \tilde{\Omega}(1)$), hence is useful in achieving near-optimal sample complexity.

Algorithm FIND-ELEMENT
Input: Parameter ϵ, distribution p.
Parameters: $m = 16/\epsilon$, $\beta_j = j\epsilon/8$, $\alpha_j = 1/(4j\log(16/\epsilon))$.

1. Draw m independent samples $x_1, x_2 \ldots x_m$ from p.
2. Output tuples $(x_1, \beta_1, \alpha_1), (x_2, \beta_2, \alpha_2), \ldots, (x_m, \beta_m, \alpha_m)$.

Lemma 3 (Appendix B.2)**.** *For $1 \leq i \leq k$, let a_i be such that $0 \leq a_i \leq 2$. If $\sum_{i=1}^{k} p_i a_i \geq \epsilon/4$, then with probability $\geq 1/5$, at least one tuple (x, α, β) returned by* FIND-ELEMENT(ϵ, p) *satisfy the property that x is α-heavy and $a_x \geq \beta$. Furthermore it uses $16/\epsilon$ samples.*

We now use the above lemma to pick elements such that $p(i) > q(i)$. Since $||p - q||_1 \geq \epsilon$,

$$\sum_{i:p(i)\geq q(i)} (p(i) - q(i)) \geq \epsilon/2.$$

Hence

$$\sum_i p(i) \max\left(0, \frac{p(i) - q(i)}{p(i)}\right) \geq \frac{\epsilon}{2}.$$

Applying Lemma 3 with $a_i = \max\left(0, \frac{p(i)-q(i)}{p(i)}\right)$, yields

Lemma 4. *If $||p - q||_1 \geq \epsilon$, then with probability $\geq 1/5$ at least one of the tuple (i, β, α) returned by* FIND-ELEMENT(ϵ, p) *satisfies $p(i) - q(i) \geq \beta p(i)$ and i is α-heavy. Furthermore* FIND-ELEMENT *uses $16/\epsilon$ samples.*

Note that even though the above algorithm does not use distribution q, it finds i such that $p(i) - q(i) \geq \beta p(i)$ just by the properties of ℓ_1 distance. Furthermore, β_j increases with j and α_j decreases with j; thus the above lemma states that the algorithm finds an element i such that either $(p(i) - q(i))/p(i)$ is large, but may not be heavy, or $(p(i) - q(i))/p(i)$ is small, yet it belongs to one of the higher probabilities. This precise trade-off becomes important to bound the sample complexity.

2.3 Testing for near-uniform distributions

We define a distribution p to be *near-uniform* if $\max_i p(i) \leq 2\min_i p(i)$. Recall that we need to find a distinguishing element and a distinguishing set. As we show, for near-uniform distributions, there are singleton distinguishing sets and hence are easy to find. Using FIND-ELEMENT, we first

define a meta algorithm to test for near-uniform distributions. The inputs to the algorithm are parameter ϵ, error δ, distributions p, q and an element y such that $p(y) \geq q(y)$. Since we use NEAR-UNIFORM-IDENTITY-TEST as a subroutine later, y is given from the main algorithm. However, if we want to use NEAR-UNIFORM-IDENTITY-TEST by itself, we can find a y using FIND-ELEMENT(ϵ, p).

The algorithm uses FIND-ELEMENT to find an element x such that $q(x) - p(x) \geq \beta q(x)$. Since $p(y) \geq q(y)$ and $q(x) - p(x) \geq \beta q(x)$, running TEST-EQUAL on set $\{x, y\}$ will yield an algorithm for identity testing. The precise bounds in Lemmas 1 and 3 help us to obtain the optimal sample complexity. In particular,

Lemma 5 (Appendix B.3)**.** *If $p = q$, then* NEAR-UNIFORM-IDENTITY-TEST *returns* `same` *with probability $\geq 1 - \delta$. If p is near-uniform and $||p - q||_1 \geq \epsilon$, then* NEAR-UNIFORM-IDENTITY-TEST *returns* `diff` *with probability $\geq 1/5 - \delta$. The algorithm uses $\mathcal{O}\left(\frac{1}{\epsilon^2} \cdot \log \frac{1}{\delta\epsilon}\right)$ samples.*

Algorithm NEAR-UNIFORM-IDENTITY-TEST
Input: distance ϵ, error δ, distributions p, q, an element y such that $p(y) \geq q(y)$.

1. Run FIND-ELEMENT(ϵ, q) to obtain tuples (x_j, β_j, α_j) for $1 \leq j \leq 16/\epsilon$.
2. For every tuple (x_j, β_j, α_j), run TEST-EQUAL$(\beta_j^2/144, 6\delta/(\pi^2 j^2), p_{\{x,y\}}, q_{\{x,y\}})$.
3. Output `same` if TEST-EQUAL in previous step returns `same` for all tuples, otherwise output `diff`.

2.4 Finding a distinguishing set for general distributions

We now extend NEAR-UNIFORM-IDENTITY-TEST to general distributions. Recall that we need to find a distinguishing element and a distinguishing set.

Once we have an element such that $p(i) > q(i)$, our objective is to find a distinguishing set S such that $p(S) < q(S)$ and $p(S) \approx p(i)$. Natural candidates for such sets are combinations of elements whose probabilities $\leq p(i)$. Since p is known, we can select such sets easily. Let $G_i = \{j : j \geq i\}$. Consider the sets $H_1, H_2, \ldots$ formed by combining elements in G_i such that $p(i) \leq p(H_j) \leq 2p(i), \forall j$. We ideally would like to use one of these H_js as S, however depending on the values of $p(H_j)$ three possible scenarios arise and that constitutes the main algorithm.

We need one more definition for describing the main identity test. For any distribution p, and a partition of S into disjoint subsets $\mathcal{S} = \{S_1, S_2, ..\}$, the induced distribution $p^{\mathcal{S}}$ is a distribution over $S_1, S_2, \ldots$ such that $\forall i$, $p^{\mathcal{S}}_S(S_i) = \frac{p(S_i)}{p(S)}$.

2.5 Proposed identity test

The algorithm is a combination of tests for each possible scenarios. First it finds a set of tuples (i, β, α) such that one tuple satisfies $(p(i) - q(i))/p(i) \geq \beta$ and i is α-heavy. Then, it divides G_i into $H_1, H_2, \ldots$ such that $, \forall j$, $p(i) \leq p(H_j) \leq 2p(i)$. If $||p - q||_1 \geq \epsilon$, then there are three possible cases.

1. $p(H_j)(1-\beta/2) \leq q(H_j)$ for most js. We can randomly pick a set H_j and sample from $H_j \cup \{i\}$ and we would be able to test if $||p - q||_1 \geq \epsilon$ using $n(i)$ when sampled from $H_j \cup \{i\}$.

2. $p(H_j)(1-\beta/2) \geq q(H_j)$ for most j. Since for most j's, $p(H_j)(1-\beta/2) \geq q(H_j)$, we have $p(G_i)(1-\beta/2) \geq q(G_i)$, and since $p(G_i) \geq \alpha$, we can sample from the entire distribution and use $n(G_i)$ to test if $||p-q||_1 \geq \epsilon$.

3. For some j, $p(H_j)(1-\beta/2) \geq q(H_j)$ and for some j, $p(H_j)(1-\beta/2) \leq q(H_j)$. It can be shown that this condition implies that elements in G_i can be grouped into $H_1, H_2, \ldots$ such that induced distribution on groups is near-uniform and yet the ℓ_1 distance between the induced distributions is large. We use NEAR-UNIFORM-IDENTITY-TEST for this scenario.

The algorithm has a step corresponding to each of the above three scenarios. If $p = q$, then all three steps would output `same` with high probability, otherwise one of the steps would output `diff`. The main result of this section is to bound the sample complexity of IDENTITY-TEST

Theorem 6 (Appendix B.4)**.** *If $p = q$, then* IDENTITY-TEST *returns* `same` *with probability $\geq 1-\delta$ and if $||p-q||_1 \geq \epsilon$, then* IDENTITY-TEST *returns* `diff` *with probability $\geq 1/30$. The algorithm uses at most $N^*_{id}(k,\epsilon) \leq \Theta\left(\frac{1}{\epsilon^2} \cdot \log^2 \frac{1}{\epsilon} \cdot \log \frac{1}{\epsilon\delta}\right)$ samples.*

The proposed identity testing has different error probabilities when $p = q$ and $||p-q||_1 \geq \epsilon$. In particular, if $p = q$, the algorithm returns `same` with probability $\geq 1-\delta$ and if $||p-q||_1 \geq \epsilon$ it outputs `diff` with probability $\geq 1/30$. While the probability of success for $||p-q||_1 \geq \epsilon$ is small, it can be boosted arbitrarily close to 1, by repeating the algorithm $\mathcal{O}(\log(1/\delta))$ times and testing if more than 1/60 fraction of times the algorithm outputs `diff`. By a simple Chernoff type argument, it can be shown that for both cases $p = q$ and $||p-q||_1$, the error probability of the boosted algorithm is $\leq \delta$. Furthermore, throughout the paper we have calculated all the constants except sample complexities which we have left in $\mathcal{O}$ notation.

Algorithm IDENTITY-TEST
Input: error δ, distance ϵ an unknown distribution q, and a known distribution p.

1. Run FIND-ELEMENT (ϵ, p) to obtain tuples (x, β, α).
2. For every tuple (x, β, α):
   (a) Let $G_x = \{y : y \geq x\}$.
   (b) Partition G_x into groups $\mathcal{H} = H_1, H_2, \ldots$s.t. for each group H_j, $p(x) \leq p(H_j) \leq 2p(x)$.
   (c) Take a random sample y from $p^{\mathcal{H}}_{G_x}$ and run TEST-EQUAL$\left(\frac{\beta^2}{1800}, \frac{\epsilon\delta}{48}, p_{\{x,y\}}, q_{\{x,y\}}\right)$.
   (d) Run TEST-EQUAL$\left(\left(\frac{\alpha\beta}{5}\right)^2, \frac{\epsilon\delta}{48}, p^{\{G_x, G_x^c\}}, q^{\{G_x, G_x^c\}}\right)$.
   (e) Run NEAR-UNIFORM-IDENTITY-TEST$\left(\frac{\beta}{5}, \frac{\epsilon\delta}{48}, p^{\mathcal{H}}_{G_x}, q^{\mathcal{H}}_{G_x}\right)$.
3. Output `diff` if any of the above tests returns `diff` for any tuple, otherwise output `same`.

3 Closeness testing

Recall that in closeness testing, both p and q are unknown and we test if $p = q$ or $||p-q||_1 \geq \epsilon$ using samples. First we relate identity testing to closeness testing.

Identity testing had two parts: finding a distinguishing element i and a distinguishing set S. The algorithm we used to generate i did not use any a priori knowledge of the distribution. Hence it carries over to closeness testing easily. The main difficulty of extending identity testing to closeness testing is to find a distinguishing set. Recall that in identity testing, we ordered elements such that their probabilities are decreasing and considered set $G_i = \{j : j \geq i\}$ to find a distinguishing set. G_i was known in identity testing, however in closeness testing, it is unknown and is difficult to find.

The rest of the section is organized as follows: We first outline a method of identifying a distinguishing set by sampling at a certain frequency (which is unknown). We then formalize finding a distinguishing element and then show how one can use a binary search to find the sampling frequency and a distinguishing set. We finally describe our main closeness test, which requires few additional techniques to handle some *special cases*.

3.1 Outline for finding a distinguishing set

Recall that in identity testing, we ordered elements such that their probabilities are decreasing and considered $G_i = \{j : j \geq i\}$. We then used a subset of $S \subset G_i$ such that $p(S) \approx p(i)$ as the distinguishing set. However, in closeness test this is not possible as set G_i is unknown. We now outline a method of finding such a set S using random sampling without the knowledge of G_i.

Without loss of generality, assume that elements are ordered such that $p(1)+q(1) \geq p(2)+q(2) \geq \ldots \geq p(k)+q(k)$. The algorithm does not use this fact and the assumption is for the ease of proof notation. Let $G_i = \{j : j \geq i\}$ under this ordering (G_i serves same purpose as G_i for identity testing, however is symmetric with respect to p, q and hence easy to handle compared to that of identity testing). Furthermore, for simplicity in the rest of the section, assume that $p(i) > q(i)$ and $p(G_i) \leq q(G_i)$. Suppose we come up with a scheme that finds subset S of G_i such that $p(S) \approx p(i)$ and $p(S) < q(S)$, then as in Identity-test, we can use that scheme together with Test-equal on $S \cup \{i\}$ to differentiate between $p = q$ and $||p - q||_1 \geq \epsilon$.

The main challenge of the algorithm is to find a distinguishing subset of G_i. Let $r = (p+q)/2$, *i.e.*, $r(j) = (p(j) + q(j))/2 \ \forall 1 \leq j \leq k$. Suppose we know $r_0 = \frac{r(i)}{r(G_i)}$. Consider a set S formed by including each element j independently with probability r_0. Thus the probability of that set can be written as

$$p(S) = \sum_{j=1}^{k} \mathbb{I}_{j \in S} p(j),$$

where $\mathbb{I}_{j \in S}$ is the indicator random variable for $j \in S$. In any such set S, there might be elements that are not from G_i. We can prune these elements (refer to them as j') by sampling from the distribution $p_{\{j,j'\}}$ and testing if j' appeared more than j. Precise probabilistic arguments are given later. Suppose we remove all elements in S that are not in G_i. Then,

$$p(S) = \sum_{j \in G_i} \mathbb{I}_{j \in S} p(j).$$

Since $\Pr(\mathbb{I}_{j \in S} = 1) = r_0$,

$$\mathbb{E}[p(S)] = \sum_{j \in G_i} \mathbb{E}[\mathbb{I}_{j \in S}] p(j) = r_0 \sum_{j \in G_i} p(j) = \frac{r(i)}{r(G_i)} \cdot p(G_i).$$

Similarly one can show that $\mathbb{E}[q(S)] = \frac{r(i)}{r(G_i)} \cdot q(G_i)$. Thus $\mathbb{E}[p(S)] < \mathbb{E}[q(S)]$ and $\mathbb{E}[p(S)] + \mathbb{E}[q(S)] = p(i) + q(i)$. Note that for efficiently using Test-equal, we not only need $p(i) > q(i)$ and $\mathbb{E}[p(S)] <$

$\mathbb{E}[q(S)]$, but we the chi-squared distance needs to be large. It can be shown that this condition is same as stating $p(S) + q(S) \approx p(i) + q(i)$ is necessary and hence $\mathbb{E}[p(S)] + \mathbb{E}[q(S)] = p(i) + q(i)$ is useful.

Thus in expectation S is a good candidate for distinguishing set. Hence if we take samples from $S \cup \{i\}$ and compare $p(S), p(i)$ and $q(S), q(i)$, we can test if $p = q$ or $||p - q||_1 \geq \epsilon$.

We therefore have to find an i such that $p(i) > q(i)$ and $p(G_i) < q(G_i)$, estimate $r(i)/r(G_i)$ and convert the above expectation argument to a probabilistic one. While the calculations and analysis in expectation seem natural, judiciously analyzing the success probability of these events takes a fair amount of effort. Furthermore, note that given a conditional sampling access to p and q, one can generate a conditional sample from r, by selecting p or q independently with probability $1/2$ and then obtaining a conditional sample from the selected distribution.

3.2 Finding a distinguishing element i

We now show that using an algorithm similar to FIND-ELEMENT, we can find an i such that $(p(i) > q(i)$ and $p(G_i) \leq q(G_i))$ or $(p(i) < q(i)$ and $p(G_i) > q(G_i))$. To quantify the above statement we need the following definition. We define β-approximability as

Definition 7. *For a pair of distributions p and q, element i is β-approximable, if*

$$\left| \frac{p(i) - q(i)}{p(i) + q(i)} - \frac{p(G_i) - q(G_i)}{p(G_i) + q(G_i)} \right| \geq \beta.$$

As we show later, it is sufficient to consider β-approximable elements instead of elements with $p(i) > q(i)$ and $p(G_i) \leq q(G_i)$. Thus the first step of our algorithm is to find β-approximable elements. To this end, we show that

Lemma 8 (Appendix C.1)**.** *If $||p - q||_1 \geq \epsilon$, then*

$$\sum_i \frac{p(i) + q(i)}{2} \left| \frac{p(i) - q(i)}{p(i) + q(i)} - \frac{p(G_i) - q(G_i)}{p(G_i) + q(G_i)} \right| \geq \frac{\epsilon}{4}.$$

Hence if we use FIND-ELEMENT for the distribution $r = (p+q)/2$, then one of the tuples would be β_j-approximable for some β_j. Note that with $a_i = \left| \frac{p(i)-q(i)}{p(i)+q(i)} - \frac{p(G_i)-q(G_i)}{p(G_i)+q(G_i)} \right|$, $0 \leq a_i \leq 2$ and $\sum_{i=1}^{k} r(i) a_i \geq \epsilon/4$. By Lemma 3, FIND-ELEMENT outputs a tuple (i, α, β) such that i that is α-heavy and β-approximable. Note that although we obtain i and guarantees on G_i, the algorithm does not find G_i.

Lemma 9. *With probability $\geq 1/5$, of the tuples returned by* FIND-ELEMENT(ϵ, r) *there exist at least one tuple that is both α-heavy and β-approximable.*

3.3 Approximating $\frac{r(i)}{r(G_i)}$ via binary search

Our next goal is to estimate $r_0 = \frac{r(i)}{r(G_i)}$ using samples. It can be easily shown that it is sufficient to know $\frac{r(i)}{r(G_i)}$ up-to a multiplicative factor, say γ (we later choose $\gamma = \Theta(\log \log \log k)$). Furthermore by the definition of G_i, $r(G_i) \geq r(i)$ and $r(G_i) = \sum_{j \geq i} r(j) \leq \sum_{j \geq i} r(i) \leq k r(i)$. Therefore,

$$\frac{1}{k} \leq \frac{r(i)}{r(G_i)} \leq 1,$$

and $\log k \geq -\log \frac{r(i)}{r(G_i)} \geq 0$. Approximating $\frac{r(i)}{r(G_i)}$ up-to a multiplicative factor γ is the same as approximating $\log \frac{r(i)}{r(G_i)}$ up-to an additive factor $\log \gamma$. We can thus run our algorithm for $\frac{r(i)}{r(G_i)}$ corresponding to each value of $\{0, \log \gamma, 2\log \gamma, 3\log \gamma, \ldots, \log k\}$, and if $||p-q||_1 \geq \epsilon$, at least for one value of $\frac{r(i)}{r(G_i)}$ we output `diff`. Using carefully chosen thresholds we can also ensure that if $p = q$, the algorithm outputs `same` always. The sample complexity for the above algorithm is $\frac{\log k}{\log \gamma} \approx \tilde{\Theta}(\log k)$ times the complexity when we know $\frac{r(i)}{r(G_i)}$. We improve the sample complexity by using a better search algorithm over $\{0, \log \gamma, 2\log \gamma, 3\log \gamma, \ldots, \log k\}$. We develop a comparator (step 4 in BINARY-SEARCH) with the following property: if our guess value $r_{\text{guess}} \geq \gamma \frac{r(i)}{r(G_i)}$ it outputs `heavy` and if $r_{\text{guess}} \leq \frac{1}{\gamma} \cdot \frac{r(i)}{r(G_i)}$ it outputs `light`. Using such a comparator, we do a binary search and find the right value faster. Recall that binary search over m elements uses $\log m$ queries. For our problem $m = \log k$ and thus our sample complexity is approximately $\log \log k$ times the sample complexity of the case when we know $\frac{r(i)}{r(G_i)}$.

However, our comparator cannot identify if we have a good guess *i.e.,* if $\frac{1}{\gamma} r_{\text{guess}} \frac{r(i)}{r(G_i)} \leq r_{\text{guess}} \leq \gamma \cdot \frac{r(i)}{r(G_i)}$. Thus, our binary search instead of outputting the value of $\frac{r(i)}{r(G_i)}$ up-to some approximation factor γ, finds a set of candidates $r^1_{\text{guess}}, r^2_{\text{guess}}, \ldots$ such that at least one of the r^j_{guess}s satisfies

$$\frac{1}{\gamma} \frac{r(i)}{r(G_i)} \leq r^j_{\text{guess}} \leq \gamma \frac{r(i)}{r(G_i)}.$$

Hence, for each value of r_{guess} we assume that $r_{\text{guess}} \approx r(i)/r(G_i)$, and run the closeness test. At least for one value of r_{guess} we would be correct. The algorithm is given in BINARY-SEARCH.

The algorithm PRUNE-SET removes all elements of probability $\geq 4r(i)$, yet does not remove any element of probability $\leq r(i)$. Since after pruning S only contains elements of probability $\leq 4r(i)$, we show that at some point of the $\log \log k$ steps, the algorithm encounters $r_{\text{guess}} \approx \frac{r(i)}{r(G_i)}$.

Algorithm PRUNE-SET
Input: S, ϵ, i, α, m, and γ.
Parameters: $\delta' = \frac{\delta}{40m \log \log k}$, $n_1 = \mathcal{O}\left(\left(\log \frac{\gamma}{\delta' \alpha\beta}\right) \cdot \left(\frac{\gamma}{\alpha\beta} \log \frac{\gamma}{\alpha\beta} + \log \frac{1}{\delta'} \log \log \frac{1}{\delta'}\right)\right)$, $n_2 = \mathcal{O}(\log \log \log k + \log \frac{1}{\epsilon\delta})$.
Repeat n_1 times:
Obtain a sample j from r_S and sample n_2 times from $r_{\{j,i\}}$. If $n(j) \geq 3n_2/4$, remove j from set S.

Algorithm BINARY-SEARCH
Input: Tuple (i, β, α).
Parameters: $\gamma = 1000 \log \frac{\log \log k}{\delta\epsilon}$, $n_3 = \mathcal{O}\left(\gamma^2 \log \frac{\log \log k}{\delta}\right)$.
Initialize $\log r_{\text{guess}} = -\log \sqrt{k}$. Set low $= -\log k$ and high $= 0$. Do $\log \log k$ times:

1. Create a set S by independently keeping elements $\{1, 2, \ldots, k\} \setminus \{i\}$ each w.p. r_{guess}.
2. Prune S using PRUNE-SET$(S, \epsilon, i, \alpha, 1, \gamma)$.
3. Run ASSISTED-CLOSENESS-TEST$(r_{\text{guess}}, (i, \beta, \alpha), \gamma, \epsilon, \delta)$.
4. Obtain n_3 samples from $S \cup \{i\}$. If $n(i) < \frac{5n_3}{\gamma}$, then output `heavy`, else output `light`.
   (a) If output is `heavy`, update high $= \log r_{\text{guess}}$ and $\log r_{\text{guess}} = (\log r_{\text{guess}} + \text{low})/2$.
   (b) If output is `light`, update low $= \log r_{\text{guess}}$ and $\log r_{\text{guess}} = (\log r_{\text{guess}} + \text{high})/2$.
5. If any of the ASSISTED-CLOSENESS-TESTs return `diff`, then output `diff`.

Lemma 10 (Appendix C.2)**.** *If i is α-heavy and β-approximable, then the algorithm* BINARY-SEARCH*, with probability $\geq 1 - \delta$, reaches r_{guess} such that*

$$\frac{r(i)}{\gamma} = \frac{r(G_i)}{\gamma} \cdot \frac{r(i)}{r(G_i)} \leq r_{guess} \leq \frac{\gamma}{\beta} \cdot \frac{r(i)}{r(G_i)}.$$

Note that due to technical reasons we get an additional $1/\beta$ factor in the upper bound and a factor of $r(G_i)$ in the lower bound.

3.4 Assisted closeness test

We now discuss the proposed test, which uses the above value of r_{guess}. As stated before, in expectation it would be sufficient to keep elements in the set S with probability r_{guess} and use the resulting set S to test for closeness. However, there are two caveats. Firstly, PRUNE-SET can remove only elements which are bigger than $4(i)$, while we can reduce the factor 4 to any number > 1, but we can never reduce it to 1 as if there is an element with probability $1 + \delta'$ for sufficiently small δ', that element is almost indistinguishable from an element with probability $1 - \delta'$. Thus we need a way of ensuring that elements with probability $> r(i)$ and $\leq 4r(i)$ do not affect the concentration inequalities.

Secondly, since we have an approximate value of $r(i)/r(G_i)$, the probability that required quantities concentrate is small and we have to repeat it many times to obtain a higher probability of success. Our algorithm address both these issues and is given below:

The algorithm picks m sets and prunes them to ensure that none of the elements has probability $\geq 4r(i)$ and considers two possibilities: there exist many elements j such that $j \notin G_i$ and

$$\left| \frac{p(i) - q(i)}{r(i)} - \frac{p(j) - q(j)}{r(j)} \right| \geq \beta'' \ (\beta'' \text{ determined later}),$$

or the number of such elements is small. If it is the first case, the algorithm finds such an element j and performs TEST-EQUAL over set $\{i, j\}$. Otherwise, we show that $r(S) \approx r(i)$, it concentrates,

and with high probability

$$\left|\frac{p(i)-q(i)}{r(i)} - \frac{p(S)-q(S)}{r(S)}\right| \geq \beta'' \text{ } (\beta'' \text{ determined later}),$$

and thus one can sample from $S \cup \{i\}$ and use $n(i)$ to test closeness.

To conclude, the proposed Closeness-test uses Find-element to find a distinguishing element i. It then runs Binary-search to approximate $r(i)/r(G_i)$. However since the search does not identify if it has found a good estimate of $r(i)/r(G_i)$, for each estimate it runs Assisted-closeness-test which uses the distinguishing element i and the estimate of $r(i)/r(G_i)$. The main result in this section is the sample complexity of our proposed Closeness-test.

Theorem 11 (Appendix C.3)**.** *If $p = q$, then* Closeness-test *returns* `same` *with probability $\geq 1-\delta$ and if $||p-q||_1 \geq \epsilon$, then* Closeness-test *returns* `diff` *with probability $\geq 1/30$. The algorithm uses $N^*_{cl}(k,\epsilon) \leq \widetilde{\mathcal{O}}\left(\frac{\log\log k}{\epsilon^5}\right)$ samples.*

As stated in the previous section, by repeating and taking a majority, the success probability can be boosted arbitrarily close to 1. Note that none of the constants or the error probabilities have been optimized. Constants for all the parameters except the sample complexities n_1, n_2, n_3, and n_4 have been given.

Algorithm Closeness-test
Input: ϵ, oracles p, q.

1. Generate a set of tuples using Find-element(ϵ, r).
2. For every tuple (i, α, β), run Binary-search(i, β, α).
3. If any of the Binary-search returned `diff` output `diff` otherwise output `same`.

Algorithm Assisted-closeness-test
Input: r_{guess}, tuple (i, β, α), γ, ϵ, and δ.
Parameters: $\beta'' = \frac{\alpha\beta}{128\gamma \log \frac{128\gamma}{\beta^2}}$, $m = \frac{4096\gamma}{\alpha\beta^2}$, $n_4 = O(\gamma/(\alpha\beta))$, and $\delta' = \frac{\epsilon\delta}{32m(n_4+1)\log\log k}$.

1. Create $S_1, S_2, \ldots S_m$ independently by keeping elements $\{1, 2, \ldots, k\} \setminus \{i\}$ each w.p. r_{guess}.
2. Run Prune-set$(S_\ell, \epsilon, i, \alpha, m, \gamma)$ for $1 \leq \ell \leq m$
3. For each set S do:
   (a) Take n_4 samples from $r_{S\cup\{i\}}$ and for all seen elements j, run Test-equal $\left((\beta'')^2/25, \delta', p_{\{i,j\}}, q_{\{i,j\}}\right)$.
   (b) Let $\mathcal{S} = \{\{i\}, S\}$. Run Test-equal$\left(\frac{\alpha^3\beta^3}{2^{23}\gamma^2\log^3\frac{128}{\gamma\beta^2}}, \delta', p^{\mathcal{S}}_{S_1\cup\{i\}}, q^{\mathcal{S}}_{S_1\cup\{i\}}\right)$.
4. If any of the above tests return `diff`, output `diff`.

4 Acknowledgements

We thank Jayadev Acharya, Clément Canonne, Sreechakra Goparaju, and Himanshu Tyagi for useful suggestions and discussions.

References

J. Acharya and C. Daskalakis. Testing poisson binomial distributions. In *Proceedings of the Twenty-Sixth Annual ACM-SIAM Symposium on Discrete Algorithms, SODA 2015, San Diego, CA, USA, January 4-6, 2015*, pages 1829–1840, 2015. 1.2

J. Acharya, H. Das, A. Jafarpour, A. Orlitsky, S. Pan, and A. T. Suresh. Competitive classification and closeness testing. In *Proceedings of the 25th Annual Conference on Learning Theory (COLT)*, pages 22.1–22.18, 2012. 1.2, A, 13

J. Acharya, C. L. Canonne, and G. Kamath. A chasm between identity and equivalence testing with conditional queries. *CoRR*, abs/1411.7346, 2014a. 1.2, 1.3

J. Acharya, A. Jafarpour, A. Orlitksy, and A. T. Suresh. Sublinear algorithms for outlier detection and generalized closeness testing. In *Proceedings of the 2014 IEEE International Symposium on Information Theory (ISIT)*, 2014b. 1.2

M. Balcan, E. Blais, A. Blum, and L. Yang. Active property testing. In *53rd Annual IEEE Symposium on Foundations of Computer Science, FOCS 2012, New Brunswick, NJ, USA, October 20-23, 2012*, pages 21–30, 2012. 1.2

T. Batu, L. Fortnow, R. Rubinfeld, W. D. Smith, and P. White. Testing that distributions are close. In *Annual Symposium on Foundations of Computer Science (FOCS)*, pages 259–269, 2000. 1.1

T. Batu, L. Fortnow, E. Fischer, R. Kumar, R. Rubinfeld, and P. White. Testing random variables for independence and identity. In *42nd Annual Symposium on Foundations of Computer Science, FOCS 2001, 14-17 October 2001, Las Vegas, Nevada, USA*, pages 442–451, 2001. 1.1

T. Batu, R. Kumar, and R. Rubinfeld. Sublinear algorithms for testing monotone and unimodal distributions. In *Proceedings of the 36th Annual ACM Symposium on Theory of Computing, Chicago, IL, USA, June 13-16, 2004*, pages 381–390, 2004. 1.2

C. L. Canonne, D. Ron, and R. A. Servedio. Testing equivalence between distributions using conditional samples. In *Proceedings of the Twenty-Fifth Annual ACM-SIAM Symposium on Discrete Algorithms, SODA 2014, Portland, Oregon, USA, January 5-7, 2014*, pages 1174–1192, 2014. 1.2

S. Chakraborty, E. Fischer, Y. Goldhirsh, and A. Matsliah. On the power of conditional samples in distribution testing. In *Innovations in Theoretical Computer Science, ITCS '13, Berkeley, CA, USA, January 9-12, 2013*, pages 561–580, 2013. 1.2

S. O. Chan, I. Diakonikolas, R. A. Servedio, and X. Sun. Efficient density estimation via piecewise polynomial approximation. In *Symposium on Theory of Computing, STOC 2014, New York, NY, USA, May 31 - June 03, 2014*, pages 604–613, 2014a. 1.2

S. O. Chan, I. Diakonikolas, P. Valiant, and G. Valiant. Optimal algorithms for testing closeness of discrete distributions. In *Symposium on Discrete Algorithms (SODA)*, 2014b. 1.1

C. Daskalakis, I. Diakonikolas, R. A. Servedio, G. Valiant, and P. Valiant. Testing k-modal distributions: Optimal algorithms via reductions. In *Proceedings of the Twenty-Fourth Annual ACM-SIAM Symposium on Discrete Algorithms, SODA 2013, New Orleans, Louisiana, USA, January 6-8, 2013*, pages 1833–1852, 2013. 1.2

I. Diakonikolas, D. M. Kane, and V. Nikishkin. Testing identity of structured distributions. In *Proceedings of the Twenty-Sixth Annual ACM-SIAM Symposium on Discrete Algorithms, SODA 2015, San Diego, CA, USA, January 4-6, 2015*, pages 1841–1854, 2015. 1.2

E. Fisher. Distinguishing distributions with conditional samples. *Bertinoro 2014*, 2014. URL http://sublinear.info/66. (document), 1.2, 1.3

O. Goldreich and D. Ron. On testing expansion in bounded-degree graphs. *Electronic Colloquium on Computational Complexity (ECCC)*, 7(20), 2000. 1.1

S. Guha, A. McGregor, and S. Venkatasubramanian. Sublinear estimation of entropy and information distances. *ACM Transactions on Algorithms*, 5(4), 2009. 1.2

R. Levi, D. Ron, and R. Rubinfeld. Testing properties of collections of distributions. *Theory of Computing*, 9:295–347, 2013. 1.2

L. Paninski. A coincidence-based test for uniformity given very sparsely sampled discrete data. *IEEE Transactions on Information Theory*, 54(10):4750–4755, 2008. 1.1

R. Rubinfeld and R. A. Servedio. Testing monotone high-dimensional distributions. *Random Struct. Algorithms*, 34(1):24–44, 2009. 1.2

J. Unnikrishnan. On optimal two sample homogeneity tests for finite alphabets. In *Proceedings of the 2012 IEEE International Symposium on Information Theory (ISIT)*, pages 2027–2031, 2012. 1.1

G. Valiant and P. Valiant. Instance-by-instance optimal identity testing. *Electronic Colloquium on Computational Complexity (ECCC)*, 20:111, 2013. 1.1, 1.2

P. Valiant. Testing symmetric properties of distributions. *SIAM Journal on Computing*, 40(6): 1927–1968, December 2011. ISSN 0097-5397. 1.1

B. Waggoner. L_p testing and learning of discrete distributions. In *Proceedings of the 2015 Conference on Innovations in Theoretical Computer Science, ITCS 2015, Rehovot, Israel, January 11-13, 2015*, pages 347–356, 2015. 1.2

J. Ziv. On classification with empirically observed statistics and universal data compression. *IEEE Transactions on Information Theory*, 34(2):278–286, 1988. 1.1

A Tools

We use the following variation of the Chernoff bound.

Lemma 12 (Chernoff bound)**.** *If $X_1, X_2 \dots X_n$ are distributed according to Bernoulli p, then*

$$\Pr\left(\frac{\sum_{i=1}^{n} X_i}{n} - p > \delta\right) \leq e^{-2n\delta^2},$$
$$\Pr\left(\frac{\sum_{i=1}^{n} X_i}{n} - p < -\delta\right) \leq e^{-2n\delta^2}.$$

The following lemma follows from a bound in Acharya et al. [2012] and the fact that $y^5 e^{-y}$ is bounded for all non-negative values of y.

Lemma 13 (Acharya et al. [2012])**.** *For two independent Poisson random variables μ and μ' with means λ and λ' respectively,*

$$E\left(\frac{(\mu-\mu')^2 - \mu - \mu'}{\mu+\mu'-1}\right) = \frac{(\lambda-\lambda')^2}{\lambda+\lambda'}\left(1 - e^{-\lambda-\lambda'}\right),$$
$$\mathrm{Var}\left(\frac{(\mu-\mu')^2 - \mu - \mu'}{\mu+\mu'-1}\right) \leq 4\frac{(\lambda-\lambda')^2}{\lambda+\lambda'} + c^2,$$

where c is a universal constant.

B Identity testing proofs

B.1 Proof of Lemma 1

Let $t = \frac{(n_1-n_2)^2-n_1-n_2}{n_1+n_2-1} + \frac{(n_1-n_2)^2-n_1-n_2}{n'+n''-n_1-n_2-1}$. Since we are using $\mathrm{poi}(n)$ samples, $n_1, n_2, n'-n_1, n''-n_2$ are all independent Poisson distributions with means $np, nq, n(1-p)$, and $n(1-q)$ respectively. Suppose the underlying hypothesis is $p = q$. By Lemma 13, $\mathbb{E}[t] = 0$ and since , $n_1, n_2, n'-n_1, n''-n_2$ are all independent Poisson distributions, variance of t is the sum of variances of each term and hence $\mathrm{Var}(t) \leq 2c^2$ for some universal constant c. Thus by Chebyshev's inequality

$$\Pr(t \geq n\epsilon/2) \leq \frac{8c^2}{(n\epsilon)^2} \leq \frac{1}{3}.$$

Hence by the Chernoff bound 12, after $18\log\frac{1}{\delta}$ repetitions probability that the majority of outputs is `same` is $\geq 1-\delta$. Suppose the underlying hypothesis is $\frac{(p-q)^2}{(p+q)(2-p-q)} > \epsilon$. Then by Lemma 13

$$\begin{aligned}
\mathbb{E}[t] &= \frac{n(p-q)^2}{p+q}\left(1 - e^{-n(p+q)}\right) + \frac{n(p-q)^2}{2-p-q}\left(1 - e^{-n(2-p-q)}\right) \\
&\overset{(a)}{\geq} \frac{n(p-q)^2}{p+q}\left(1 - e^{-n\epsilon}\right) + \frac{n(p-q)^2}{2-p-q}\left(1 - e^{-n\epsilon}\right) \\
&\geq \frac{2n(p-q)^2}{(p+q)(2-p-q)}\left(1 - e^{-n\epsilon}\right) \\
&\overset{(b)}{\geq} \frac{n(p-q)^2}{(p+q)(2-p-q)}.
\end{aligned}$$

(a) from the fact that $p+q \geq (p+q)\frac{(p-q)^2}{(p+q)^2} \geq \frac{(p-q)^2}{p+q} \geq \epsilon$ and similarly $2-p-q \geq \epsilon$. (b) follows the fact that $n\epsilon \geq 10$. Similarly the variance is

$$\mathrm{Var}(t) \leq c^2 + \frac{4n(p-q)^2}{p+q} + c^2 + \frac{4n(p-q)^2}{2-p-q} = 2c^2 + \frac{8n(p-q)^2}{(p+q)(2-p-q)}.$$

Thus again by Chebyshev's inequality,

$$\Pr(t \leq n\epsilon/2) \leq \frac{2c^2 + \frac{8n(p-q)^2}{(p+q)(2-p-q)}}{(\frac{n(p-q)^2}{(p+q)(2-p-q)} - \frac{n\epsilon}{2})^2} \leq \frac{32c^2}{(n\epsilon)^2} + \frac{32}{n\epsilon} \leq \frac{1}{3}.$$

The last inequality follows when $n \geq \frac{\max(192,20c)}{\epsilon}$. The lemma follows by the Chernoff bound argument as before.

B.2 Proof of Lemma 3

Let A_j be the event that $a_{x_j} \geq \beta_j$ and x_j is α_j-heavy. Since we choose each tuple j independently at time j, events A_js are independent. Hence,

$$\begin{aligned}
\Pr(\cup A_j) &= 1 - \Pr(\cap A_j^c) \\
&= 1 - \prod_{j=1}^{m} \Pr(A_j^c) \\
&= 1 - \prod_{j=1}^{m} (1 - \Pr(A_j)) \\
&\geq 1 - e^{-\sum_{j=1}^{m} \Pr(A_j)}.
\end{aligned}$$

Let $B_j = \{i : a_i \geq \beta_j\}$. Since all elements in B_j count towards A_j except the last α_j part, $\Pr(A_j) \geq p(B_j) - \alpha_j$. Thus

$$\begin{aligned}
\sum_{j=1}^{m} \Pr(A_j) &\geq \sum_{j=1}^{m} p(B_j) - \sum_{j=1}^{m} \alpha_j \\
&\geq \sum_{j=1}^{m} p(B_j) - \frac{\log m}{4 \log 16/\epsilon} \\
&\geq \sum_{j=1}^{m} p(B_j) - 1/4.
\end{aligned}$$

We now show that $\sum_{j=1}^{m} p(B_j) \geq 1/2$, thus proving that $\sum_{j=1}^{m} \Pr(A_j) \geq 1/4$ and $\Pr(\cup_j A_j) \geq 1/5$. Since $\sum_{i=1}^{k} p(i)a_i \geq \epsilon/4$,

$$\begin{aligned}
\sum_{i=1}^{m} p(i)a_i &= \sum_{i:a_i \geq \epsilon/8} p(i)a_i + \sum_{i:a_i < \epsilon/8} p(i)a_i \\
&\leq \sum_{i:a_i \geq \epsilon/8} p(i)a_i + \epsilon/8.
\end{aligned}$$

Thus

$$\sum_{i:a_i \geq \epsilon/8} p(i) a_i \geq \epsilon/8,$$

and

$$\sum_{i:a_i \geq \epsilon/8} p(i) \left\lfloor \frac{8a_i}{\epsilon} \right\rfloor \geq 1/2.$$

In $\sum_{j=1}^{m} p(B_j)$, each $p(i)$ is counted exactly $\left\lfloor \frac{8a_i}{\epsilon} \right\rfloor$ times, thus

$$\sum_{i=1}^{m} p(B_j) \geq \frac{1}{2}.$$

B.3 Proof of Lemma 5

The proof uses the following auxiliary lemma. Let $p_{i,j} \stackrel{\text{def}}{=} p_{\{i,j\}}(i) = \frac{p(i)}{p(i)+p(j)}$ denote the probability of i under conditional sampling.

Lemma 14. *If*

$$\left| \frac{p(i) - q(i)}{p(i) + q(i)} - \frac{p(j) - q(j)}{p(j) + q(j)} \right| \geq \epsilon, \tag{1}$$

then

$$\begin{aligned}
&\frac{(p_{i,j} - q_{i,j})^2}{(p_{i,j} + q_{i,j})(2 - p_{i,j} - q_{i,j})} \\
&\geq \frac{\epsilon^2 (p(i) + q(i))^2 (p(j) + q(j))^2}{4[p(i)(q(i) + q(j)) + q(i)(p(i) + p(j))][p(j)(q(i) + q(j)) + q(j)(p(i) + p(j))]} \\
&\geq \epsilon^2 \frac{(p(i) + q(i))(p(j) + q(j))}{4(p(i) + q(i) + p(j) + q(j))^2}.
\end{aligned}$$

Proof. Let $s(i) = p(i) + q(i)$ and $s(j) = p(j) + q(j)$. Upon expanding,

$$\left| \frac{p(i) - q(i)}{p(i) + q(i)} - \frac{p(j) - q(j)}{p(j) + q(j)} \right| = 2 \left| \frac{p(i)q(j) - p(j)q(i)}{s(i)s(j)} \right|. \tag{2}$$

Furthermore, $p_{i,j} = \frac{p(i)}{p(i)+p(j)}$ and similarly $q_{i,j} = \frac{q(i)}{q(i)+q(j)}$. Hence,

$$\begin{aligned}
&\frac{(p_{i,j} - q_{i,j})^2}{(p_{i,j} + q_{i,j})(2 - p_{i,j} - q_{i,j})} \\
&= \frac{(p(i)q(j) - q(i)p(j))^2}{[p(i)(q(i) + q(j)) + q(i)(p(i) + p(j))][p(j)(q(i) + q(j)) + q(j)(p(i) + p(j))]} \\
&\stackrel{(a)}{\geq} \frac{\epsilon^2 s^2(i) s^2(j)}{4[p(i)(q(i) + q(j)) + q(i)(p(i) + p(j))][p(j)(q(i) + q(j)) + q(j)(p(i) + p(j))]} \\
&\stackrel{(b)}{\geq} \frac{\epsilon^2 s^2(i) s^2(j)}{4(s(i) + s(j))^2 s(i) s(j)} \\
&= \frac{\epsilon^2 s(i) s(j)}{4(s(i) + s(j))^2}.
\end{aligned}$$

(a) follows by Equations (1) and (2). (b) follows from $\max(p(i), q(i)) \leq s(i)$ and $\max(p(j), q(j)) \leq s(j)$. ■

Proof. (Lemma 5) We first show that if $p = q$, then NEAR-UNIFORM-IDENTITY-TEST returns same with probability $\geq 1 - \delta$. By Lemma 1, TEST-EQUAL returns error probability $\delta_j = 6\delta/(\pi^2 j^2)$ for the jth tuple and hence by the union bound, the overall error is $\leq \sum_j \delta_j \leq \delta$.

If $p \neq q$, with probability $\geq 1/5$, FIND-ELEMENT returns an element x that is α-heavy and $q(x) - p(x) \geq \beta q(x)$. For this x, y, since $p(y) \geq q(y)$,

$$\frac{q(x) - p(x)}{p(x) + q(x)} - \frac{q(y) - p(y)}{p(y) + q(y)} \geq \frac{\beta q(x)}{p(x) + q(x)}.$$

By Lemma 14 the chi-squared distance between $p_{\{x,y\}}$ and $q_{\{x,y\}}$ is lower bounded by

$$\begin{aligned}
&\geq \left(\frac{\beta q(x)}{p(x) + q(x)}\right)^2 \frac{(p(x) + q(x))^2 (p(y) + q(y))^2}{4[p(x)(q(x) + q(y)) + q(x)(p(x) + p(y))][p(y)(q(x) + q(y)) + q(y)(p(x) + p(y))]} \\
&\overset{(a)}{\geq} \frac{\beta^2 q^2(x) p^2(y)}{4[p(x)(q(x) + p(y)) + q(x)(p(x) + p(y))][p(y)(q(x) + p(y)) + p(y)(p(x) + p(y))]} \\
&\overset{(b)}{\geq} \frac{\beta^2 q^2(x) p^2(y)}{4[2p(y)q(x) + p(x)p(y) + q(x)(2p(y) + p(y))][p(y)(q(x) + 2p(x)) + p(y)(p(x) + 2p(x))]} \\
&\overset{(c)}{\geq} \frac{\beta^2 q^2(x) p^2(y)}{4[2p(y)q(x) + q(x)p(y) + q(x)(2p(y) + p(y))][p(y)(q(x) + 2q(x)) + p(y)(q(x) + 2q(x))]} \\
&\overset{(d)}{\geq} \frac{\beta^2}{144}.
\end{aligned}$$

(a) follows from the fact that $p(y) \geq q(y)$. $p(x) \leq 2p(y)$ and $p(y) \leq 2p(x)$ hence (b). $p(x) \leq q(x)$ implies (c) and (d) follows from numerical simplification. Thus by Lemma 1 algorithm returns `diff` with probability $\geq 1 - \delta$. By the union bound, the total error probability is $\leq \frac{4}{5} + \delta$. The number of samples used is $16/\epsilon$ for the first step and $\widetilde{\mathcal{O}}\left(\frac{1}{\beta_j^2} \log \frac{1}{\delta_j}\right)$ for tuple j. Hence the total number of samples used is

$$\frac{16}{\epsilon} + \sum_{j=1}^{16/\epsilon} \mathcal{O}\left(\frac{1}{\beta_j^2} \log \frac{1}{\delta_j}\right) = \mathcal{O}\left(\frac{1}{\epsilon^2} \log \frac{1}{\delta\epsilon}\right).$$

■

B.4 Proof of Theorem 6

We state the theorem statement for better readability: If $p = q$, then IDENTITY-TEST returns `same` with probability $\geq 1 - \delta$ and if $||p - q||_1 \geq \epsilon$, then IDENTITY-TEST returns `diff` with probability $\geq 1/30$.

Recall that there are $\frac{16}{\epsilon}$ tuples. Also observe that all the three tests inside IDENTITY-TEST are called with error parameter $\frac{\epsilon\delta}{48}$. As a result if $p = q$, IDENTITY-TEST outputs `same` with probability $\geq 1 - \frac{\epsilon\delta}{48} \cdot 3 \cdot \frac{16}{\epsilon} = 1 - \delta$.

We now show that if $||p-q||_1 \geq \epsilon$, then the algorithm outputs `diff` with probability $\geq 1/30$. By Lemma 4, with probability $\geq 1/5$ FIND-ELEMENT returns an element x such that $p(x)-q(x) \geq \beta p(x)$ and α-heavy. Partition G_x into groups $\mathcal{H} = H_1, H_2, \ldots$ s.t. for each group H_j, $p(x) \leq p(H_j) \leq 2p(x)$ and let $p^{\mathcal{H}}_{G_x}$ and $q^{\mathcal{H}}_{G_x}$ be the corresponding induced distributions. There are three possible cases. We show that for any q, at least one of the sub-routines in IDENTITY-TEST will output `diff` with high probability.

1. $|p(G_x) - q(G_x)| \geq \frac{\alpha\beta}{5}$.
2. $|p(G_x) - q(G_x)| < \frac{\alpha\beta}{5}$ and $\left|\left|p^{\mathcal{H}}_{G_x} - q^{\mathcal{H}}_{G_x}\right|\right|_1 \geq \frac{\beta}{5}$.
3. $|p(G_x) - q(G_x)| < \frac{\alpha\beta}{5}$ and $\left|\left|p^{\mathcal{H}}_{G_x} - q^{\mathcal{H}}_{G_x}\right|\right|_1 < \frac{\beta}{5}$.

If $|p(G_x) - q(G_x)| \geq \frac{\alpha\beta}{5}$, then chi-squared distance between $p^{\{G_x, G_x^c\}}$ and $q^{\{G_x, G_x^c\}}$ is $\geq \left(\frac{\alpha\beta}{5}\right)^2$ and hence TEST-EQUAL$\left((\alpha\beta/5)^2, \frac{\epsilon\delta}{48}, p^{\{G_x, G_x^c\}}, q^{\{G_x, G_x^c\}}\right)$ (step 2c) outputs `diff` with probability $> 1 - \frac{\epsilon\delta}{48}$.

If $|p(G_x) - q(G_x)| < \frac{\alpha\beta}{5}$ and $\left|\left|p^{\mathcal{H}}_{G_x} - q^{\mathcal{H}}_{G_x}\right|\right|_1 \geq \frac{\beta}{5}$, then by Lemma 5 NEAR-UNIFORM-IDENTITY-TEST$(\frac{\beta}{5}, \frac{\epsilon\delta}{48}, p^{\mathcal{H}}_{G_x}, q^{\mathcal{H}}_{G_x})$ outputs `diff` with probability $> \frac{1}{5} - \frac{\epsilon\delta}{48} > \frac{1}{6}$.

If $|p(G_x) - q(G_x)| < \frac{\alpha\beta}{5}$ and $\left|\left|p^{\mathcal{H}}_{G_x} - q^{\mathcal{H}}_{G_x}\right|\right|_1 < \frac{\beta}{5}$,

$$
\begin{aligned}
\sum_{y\in\mathcal{H}} p^{\mathcal{H}}_{G_x}(y)\mathbb{I}\left[\frac{p(y)-q(y)}{p(y)} > \frac{4}{5}\beta\right] &\leq \frac{1}{p(G_x)}\sum_{y\in\mathcal{H}} p(y)\mathbb{I}\left[p(y)-q(y) > \frac{4}{5}\beta p(y)\right] \\
&\leq \frac{5}{4\beta p(G_x)}\sum_{y\in\mathcal{H}}|p(y)-q(y)| \\
&= \frac{5}{4\beta}\sum_{y\in\mathcal{H}}\left|\frac{p(y)}{p(G_x)} - \frac{q(y)}{p(G_x)} + \frac{q(y)}{q(G_x)} - \frac{q(y)}{q(G_x)}\right| \\
&\overset{(a)}{\leq} \frac{5}{4\beta}\left(\sum_{y\in\mathcal{H}}\left|\frac{p(y)}{p(G_x)} - \frac{q(y)}{q(G_x)}\right| + \sum_{y\in\mathcal{H}} q(y)\left|\frac{1}{p(G_x)} - \frac{1}{q(G_x)}\right|\right) \\
&\overset{(b)}{\leq} \frac{5}{4\beta}\left(\frac{\beta}{5} + q(G_x)\frac{|p(G_x)-q(G_x)|}{p(G_x)q(G_x)}\right) \\
&\overset{(c)}{\leq} \frac{5}{4\beta}\left(\frac{\beta}{5} + \frac{\beta}{5}\right) \\
&\leq \frac{1}{2}.
\end{aligned}
$$

(a) follows from triangle inequality. (b) follows from the fact that $\left|\left|p^{\mathcal{H}}_{G_x} - q^{\mathcal{H}}_{G_x}\right|\right|_1 \leq \frac{\beta}{5}$. $p(G_x) \geq \alpha$ and $p(G_x) - q(G_x) \leq \frac{\alpha\beta}{5}$ and hence (c). Therefore, for a random sample y from $p^{\mathcal{H}}_{G_x}$, with probability $\geq 1/2$, $\frac{p(y)-q(y)}{p(y)} \leq \frac{4\beta}{5}$. Let $\frac{q(y)}{p(y)} = \beta' \geq 1 - \frac{4\beta}{5}$ and furthermore $\frac{q(x)}{p(x)} = \beta'' \leq 1-\beta$. Hence $\beta' - \beta'' \geq \frac{\beta}{5}$. Thus similar to the proof of Lemma 14, the chi-squared distance between $p_{\{x,y\}}$ and $q_{\{x,y\}}$ can be

lower bounded by

$$\begin{aligned}
&\geq \frac{(p(x)q(y)-q(x)p(y))^2}{[p(x)(q(x)+q(y))+q(x)(p(x)+p(y))][p(y)(q(x)+q(y))+q(y)(p(x)+p(y))]}\\
&\overset{(a_1)}{\geq} \frac{(\beta'-\beta'')^2p^2(x)p^2(y)}{[p(x)(q(x)+q(y))+q(x)(p(x)+p(y))][p(y)(q(x)+q(y))+q(y)(p(x)+p(y))]}\\
&\overset{(a_2)}{\geq} \frac{(\beta'-\beta'')^2}{\max^2(1,\beta',\beta'')}\frac{p(x)p(y)}{4(p(x)+p(y))^2}\\
&\overset{(b)}{\geq} \frac{(\beta'-\beta'')^2}{18\max^2(1,\beta',\beta'')}\\
&\overset{(c)}{\geq} \frac{\beta^2}{1800}.
\end{aligned}$$

$(a_1), (a_2)$ follow by substituting $q(x) = \beta' p(x)$ and $q(y) = \beta'' p(y)$. (b) follows from $p(x) \leq 2p(y)$ and $p(y) \leq 2p(x).\beta'' \leq 1$ and $\beta' - \beta'' \geq \frac{\beta}{5}$ and hence the RHS in (b) is minimized by $\beta' = 1 + \frac{\beta}{5}$ and $\beta'' = 1$. For these values of β', β'', $\max(1, \beta', \beta'') \leq 2$ and hence (c). Thus TEST-EQUAL outputs `diff` with probability $> 1 - \frac{\epsilon\delta}{48}$.

If $||p - q||_1 \geq \epsilon$, then by Lemma 4 step 1 picks a tuple (x, β, α) such that $p(x) - q(x) \geq p(x)\beta$ with probability at least $\frac{1}{5}$. Conditioned on this event, for the three cases discussed above the minimum probability of outputting `diff` is $\frac{1}{6}$ and every p, q falls into one of the three categories. Hence with probability $> \frac{1}{30}$ IDENTITY-TEST outputs `diff`.

We now compute the sample complexity of IDENTITY-TEST. Step 1 of the algorithm uses $16/\epsilon$ samples. For every tuple (x, β, α), step $2(c)$ of the algorithm uses $\mathcal{O}\left(\frac{1}{\beta^2}\log\frac{1}{\delta\epsilon}\right)$ samples. Summing over all tuples yields a sample complexity of

$$\sum_{j=1}^{16/\epsilon} \mathcal{O}\left(\frac{1}{\beta_j^2}\log\frac{1}{\delta_j\epsilon}\right) = \mathcal{O}\left(\frac{1}{\epsilon^2}\log\frac{1}{\delta\epsilon}\right)$$

For the different tuples TEST-EQUAL$\left(\frac{\alpha\beta}{5})^2, \frac{\epsilon\delta}{30}, p^{\{G_x, G_x^c\}}, p^{\{G_x, G_x^c\}}\right)$ can reuse samples and as $\alpha\beta = \Omega(\epsilon/(\log 1/\epsilon))$, it uses a total of $\Theta\left(\frac{1}{\epsilon^2}\log^2\frac{1}{\epsilon}\log\frac{1}{\delta\epsilon}\right)$. samples.

Furthermore, NEAR-UNIFORM-IDENTITY-TEST uses $\mathcal{O}\left(\frac{1}{\beta^2}\log\frac{1}{\epsilon\delta}\right)$ samples. Summing over all tuples, the sample complexity is $\mathcal{O}\left(\frac{1}{\epsilon^2}\log\frac{1}{\delta\epsilon}\right)$. Summing over all the three cases, the sample complexity of the algorithm is $\mathcal{O}\left(\frac{1}{\epsilon^2}\log^2\frac{1}{\epsilon}\log\frac{1}{\delta\epsilon}\right)$.

C Closeness testing proofs

C.1 Proof of Lemma 8

Recall that $G_i = \{j : j \geq i\}$. Let $r_i = \frac{p(i)+q(i)}{2}$ and $s_i = \frac{p(i)-q(i)}{2}$. We will use the following properties: $\sum_{i=1}^k r_i = 1$, $\sum_{i=1}^k s_i = 0$, and $\sum_{i=1}^k |s_i| \geq \frac{\epsilon}{2}$. We will show that

$$\sum_{i=1}^k r_i \left|\frac{s_i}{r_i} - \frac{\sum_{j=i}^k s_j}{\sum_{j=i}^k r_j}\right| \geq \frac{\epsilon}{4}.$$

We show that

$$\sum_{i=1}^{k} r_i \left| \frac{s_i}{r_i} - \frac{\sum_{j=i}^{k} s_j}{\sum_{j=i}^{k} r_j} \right| \geq \frac{|s_1| + |s_2| - |s_1 + s_2|}{2} + (r_1 + r_2) \left| \frac{s_1 + s_2}{r_1 + r_2} - \frac{\sum_{j=1}^{k} s_j}{\sum_{j=1}^{k} r_j} \right| + \sum_{i=3}^{k} r_i \left| \frac{s_i}{r_i} - \frac{\sum_{j=i}^{k} s_j}{\sum_{j=i}^{k} r_j} \right|.$$

Thus reducing the problem from k indices to $k-1$ indices with $s_1, s_2, \ldots s_k$ going to $s_1+s_2, s_3, \ldots s_k$ and $r_1, r_2, \ldots r_k$ going to $r_1+r+2, r_3, r_4, \ldots r_k$. Continuing similarly we can reduce the $k-1$ indices to $k-2$ indices with terms $s_1 + s_2 + s_3, s_4 \ldots s_k$ and $r_1 + r_2 + r_3, r_4 \ldots r_k$ and so on. Telescopically adding the sum

$$\begin{aligned} \sum_{i=1}^{k} r_i \left| \frac{s_i}{r_i} - \frac{\sum_{j=i}^{k} s_j}{\sum_{j=i}^{k} r_j} \right| &\geq \frac{|s_1| + |s_2| - |s_1 + s_2|}{2} + \frac{|s_1 + s_2| + |s_3| - |s_1 + s_2 + s_3|}{2} + \ldots \\ &= \frac{\sum_{i=1}^{k} |s_i|}{2} \geq \frac{\epsilon}{4}, \end{aligned}$$

where the last equality follows from the fact that $\sum_{i=1}^{k} s_i = 0$. To prove the required inductive step, it suffices to show

$$\begin{aligned} \sum_{i=1}^{2} r_i \left| \frac{s_i}{r_i} - \frac{\sum_{j=i}^{k} s_j}{\sum_{j=i}^{k} r_j} \right| &\geq \frac{|s_1| + |s_2| - |s_1 + s_2|}{2} + (r_1 + r_2) \left| \frac{s_1 + s_2}{r_1 + r_2} - \frac{\sum_{j=1}^{k} s_j}{\sum_{j=1}^{k} r_j} \right| \\ &\geq \frac{|s_1| + |s_2| + |s_1 + s_2|}{2}, \end{aligned}$$

where the last inequality follows from the fact that $\sum_{i=1}^{k} s_i = 0$. Rewriting the left hand side using the fact that $\sum_{i=1}^{k} s_i = 0$

$$|s_1| + r_2 \left| \frac{s_2}{r_2} + \frac{s_1}{r_2 + r_3'} \right|,$$

where $r_3' = \sum_{j=3}^{k} r_j$. Thus it suffices to show

$$|s_1| + r_2 \left| \frac{s_2}{r_2} + \frac{s_1}{r_2 + r_3'} \right| \geq \frac{|s_1| + |s_2| + |s_1 + s_2|}{2}.$$

We prove it by considering three sub-cases: s_1, s_2 have the same sign, s_1, s_2 have different signs but $|s_1| \geq |s_2|$, and s_1, s_2 have different signs but $|s_1| < |s_2|$. If s_1, s_2 have the same sign, then

$$|s_1| + r_2 \left| \frac{s_2}{r_2} + \frac{s_1}{r_2 + r_3'} \right| \geq |s_1| + r_2 \left| \frac{s_2}{r_2} \right| = |s_1| + |s_2| = \frac{|s_1| + |s_2| + |s_1 + s_2|}{2}.$$

If s_1 and s_2 have different signs and $|s_1| \geq |s_2|$, then

$$|s_1| + r_2 \left| \frac{s_2}{r_2} + \frac{s_1}{r_2 + r_3'} \right| \geq |s_1| = \frac{|s_1| + |s_1|}{2} = \frac{|s_1| + |s_2| + |s_1 + s_2|}{2}.$$

If s_1 and s_2 have different signs and $|s_1| < |s_2|$, then

$$|s_1| + r_2 \left| \frac{s_2}{r_2} + \frac{s_1}{r_2 + r_3'} \right| \geq |s_1| + r_2 \left| \frac{s_2}{r_2} + \frac{s_1}{r_2} \right| = |s_1| + |s_2 + s_1| = \frac{|s_1| + |s_2| + |s_1 + s_2|}{2}.$$

C.2 Proof of Lemma 10

We prove this lemma using several smaller sub-results. We first state a concentration result, which follows from Bernstein's inequality.

Lemma 15. *Consider a set G such that $\max_{j\in G} r(j) \le r_{\max}$. Consider set S formed by selecting each element from G independently and uniformly randomly with probability r_0, then*

$$E[r(S)] = r_0 r(G),$$

and with probability $\ge 1-2\delta$,

$$|r(S) - E[r(S)]| \le \sqrt{2r_0 r_{\max} r(G) \log\frac{1}{\delta}} + r_{\max}\log\frac{1}{\delta}.$$

Furthermore

$$E[|S|] = r_0|G|,$$

and with probability $\ge 1-2\delta$,

$$||S| - r_0|G|| \le \sqrt{2r_0|G|\log\frac{1}{\delta}} + \log\frac{1}{\delta}.$$

C.2.1 Results on Prune-set

We now show that with high probability Prune-set never removes an element with probability $\le 2r(i)$.

Lemma 16. *The probability that the algorithm removes an element j from the set S such that $r(j) \le 2r(i)$ during step 2 of* Binary-search *is $\le \delta/5$.*

Proof. If $r(j) < 2r(i)$, then $\frac{r(j)}{r(j)+r(i)} \le \frac{2}{3}$. Applying Chernoff bound,

$$\Pr\left(n(j) \ge \frac{3n_2}{4}\right) \le e^{-n_2/72}.$$

Since the algorithm uses this step no more than $\mathcal{O}(n_1 \log\log k)$ times, the total error probability is less than $\mathcal{O}(n_1 \log\log k \cdot e^{-n_2/72})$. Since n_1 is $\mathrm{poly}(\log\log\log k, \epsilon^{-1}, \log\delta^{-1})$ and $n_2 = \mathcal{O}(\log\log\log k + \log\frac{1}{\epsilon\delta})$, the error probability is $\le \delta/5$. ■

We now show that Prune-set removes all elements with probability $\ge 4r(i)$ with high probability. Recall that $\delta' = \frac{\delta}{40m\log\log k}$.

Lemma 17. *If element i is α-heavy, β-approximable and $r_{guess} \le \frac{\gamma}{\beta}\frac{r(i)}{r(G_i)}$, then* Prune-set *removes all elements such that $r(j) > 4r(i)$ during all calls of step 2 of* Binary-search *with probability $\ge 1-\frac{\delta}{5}$.*

Proof. Let $A = \{j : r(j) \le 4r(i)\}$ and $S' = S \cap A$. By Lemma 15, with probability $\ge 1-2\delta'$

$$r(S') \le r_{\text{guess}} r(A) + \sqrt{8r_{\text{guess}} r(i) r(A)\log\frac{1}{\delta'}} + 4r(i)\log\frac{1}{\delta'} \le 2r_{\text{guess}} + 8r(i)\log\frac{1}{\delta'},$$

where the last inequality follows from the identity $\sqrt{2ab} \le a+b$. Observe that $|A^c| \le \frac{1}{4r(i)}$. Let $S'' = S \setminus S'$. By Lemma 15, with probability $\ge 1-2\delta'$

$$\nu \stackrel{\text{def}}{=} |S''| \le r_{\text{guess}} \frac{1}{4r(i)} + \sqrt{2r_{\text{guess}} \frac{1}{4r(i)} \log \frac{1}{\delta'}} + \log \frac{1}{\delta'} \le \frac{r_{\text{guess}}}{2r(i)} + 2\log \frac{1}{\delta'}.$$

S has ν elements with probability $> 4r(i)$. Suppose we have observed j of these elements and removed them from S. There are $\nu - j$ of them left in S. After taking another η samples from S, the probability of not observing a $(j+1)$th heavy element is $< (r(S')/(r(S') + 4r(i)(\nu - j)))^{\eta}$. Therefore,

$$\eta_j \stackrel{\text{def}}{=} \log \frac{\nu}{\delta'} \cdot \left(1 + \frac{r(S')}{4r(i)(\nu - j)}\right) \ge \frac{\log \frac{\nu}{\delta'}}{\log\left(1 + \frac{4r(i)(\nu-j)}{r(S')}\right)}$$

samples suffice to observe an element from S'' with probability $> 1 - \frac{\delta'}{\nu}$. After observing the sample (call it j), similar to the proof of Lemma 16 it can be shown that with probability $\ge 1-\delta'$, for samples from $r_{\{j,i\}}$, $n(j) \ge 3n_2/4$ and hence j will be removed from S. Thus to remove all ν elements of probability $> 4r(i)$, we need to repeat this step

$$n_1 = \sum_{j=1}^{\nu} \eta_j = \log \frac{\nu}{\delta'} \cdot \sum_{j=1}^{\nu} \left(1 + \frac{r(S')}{4r(i)j}\right) \le \log \frac{\nu}{\delta'} \cdot \left(\nu + \frac{r(S')}{4r(i)} \log \nu\right)$$

times. Substituting $r(S')$ and ν in the RHS and simplifying we have

$$n_1 \le 4 \log \frac{\gamma}{2\delta'\alpha\beta} \left(\frac{\gamma}{2\alpha\beta} \log \frac{\gamma}{2\alpha\beta} + 2\log \frac{1}{\delta'} \log\log \frac{1}{\delta'}\right).$$

By the union bound, total error probability is $\le \delta'$. Since the number of calls to PRUNE-SET is at most $\log\log k$ during step 2 of the algorithm, the error is at most $\log\log k \cdot 2\delta' \le \delta/5$ and the lemma follows from the union bound. ■

C.2.2 Proof of Lemma 10

The proof of Lemma 10 follows from the following two sub-lemmas. In Lemma 18, we show that if $r_{\text{guess}} \ge \gamma \frac{r(i)}{r(G_i)}$ then step 4 will return `heavy`, and if $r_{\text{guess}} \le \frac{1}{\gamma} \frac{r(i)}{r(G_i)}$ hence the algorithm outputs `light` with high probability. Since we have $\log\log k$ iterations and $\frac{1}{k} \le \frac{r(i)}{r(G_i)} \le 1$, we reach $\frac{r(i)}{\gamma r(G_i)} \le r_{\text{guess}} \le \frac{\gamma r(i)}{\beta r(G_i)}$ at some point of the algorithm.

Lemma 18. *If $r_{guess} > \frac{\gamma}{\beta} \frac{r(i)}{r(G_i)}$, i is α heavy, β-approximable, and* PRUNE-SET *has removed none of the elements with probability $\le 2r(i)$, then with probability $\ge 1 - 4\delta'$, step 4 outputs* `heavy`.

Proof. Let $G_i' = G_i \setminus \{i\}$. Since i is α heavy and β-approximable, by convexity

$$\frac{r(G_i')}{r(G_i)} \left|\frac{p(i)-q(i)}{p(i)+q(i)} - \frac{p(G_i')-q(G_i')}{p(G_i')+q(G_i')}\right| \ge \left|\frac{p(i)-q(i)}{p(i)+q(i)} - \frac{p(G_i)-q(G_i)}{p(G_i)+q(G_i)}\right| \ge \beta.$$

Hence $r(G'_i) \geq \beta r(G_i)/2$. By assumption, all the elements with probability $< 2r(i)$ in set S will remain after pruning. Thus all the elements in set S from G'_i remains after pruning. Let $S' = G'_i \cap S$. By Lemma 15 with $G = G'_i$, $r_{\max} = r_i$, and $r_0 = r_{\text{guess}}$

$$\Pr\left(r(S') \leq r_{\text{guess}} r(G'_i) - \sqrt{2r_{\text{guess}} r(i) r(G'_i) \log\frac{1}{\delta'}} - r(i)\log\frac{1}{\delta'}\right) < 2\delta'. \tag{3}$$

Taking derivatives, it can be shown that the slope of the RHS of the term inside parenthesis is $r(G'_i) - \sqrt{\frac{r(i)r(G'_i)\log\frac{1}{\delta'}}{2r_{\text{guess}}}}$, which is positive for $r_{\text{guess}} \geq \frac{\gamma r(i)}{\beta r(G_i)}$. Thus the value is minimized at $r_{\text{guess}} = \frac{\gamma r(i)}{\beta r(G_i)}$ in the range $\left[\frac{\gamma r(i)}{\beta r(G_i)}, \infty\right)$ and simplifying this lower bound using values of γ, β, we get

$$r_{\text{guess}} r(G'_i) - \sqrt{2r_{\text{guess}} r(i) r(G'_i) \log\frac{1}{\delta'}} - r(i)\log\frac{1}{\delta'} \geq \frac{\gamma r(i)}{4}.$$

Since $\Pr(X < b) \leq \Pr(X < b + t)$, we have

$$\Pr\left(r(S') \leq \frac{\gamma r(i)}{4}\right) < 2\delta'.$$

Hence with probability $\geq 1 - 2\delta'$, $\frac{r(i)}{r(S)+r(i)} \leq \frac{r(i)}{r(S')} \leq \frac{4}{\gamma}$. By the Chernoff bound,

$$\Pr\left(\frac{n(i)}{n_3} > \frac{5}{\gamma}\right) \leq \Pr\left(\frac{n(i)}{n_3} > \frac{r(i)}{r(S)+r(i)} + \frac{1}{\gamma}\right) \leq e^{-2n_3/\gamma^2}.$$

Therefore, for $n_3 \geq \mathcal{O}\left(\gamma^2 \log\frac{\log\log k}{\delta}\right)$, step 3 outputs `heavy` with probability $\geq 1 - 2\delta'$. By the union bound the total error probability $\leq 4\delta'$ ■

Lemma 19. *If $r_{guess} < \frac{r(i)}{\gamma}$ and* PRUNE-SET *has removed all elements with probability $\geq 4r(i)$ and none of the elements with probability $\leq 2r(i)$, then with probability $\geq 1 - 4\delta'$, step 3, outputs* `light`.

Proof. The proof is similar to that of Lemma 18. By assumption all the elements have probability $\leq 4r(i)$. By Lemma 15,

$$\Pr\left(r(S) > r(i)\left[\sqrt{8\frac{r_{\text{guess}}}{r(i)}\log\frac{1}{\delta'}} + \frac{r_{\text{guess}}}{r(i)} + 4\log\frac{1}{\delta'}\right]\right) \leq 2\delta'.$$

Similar to the analysis after Equation (3), taking derivatives it can be shown that the RHS of the term inside parenthesis is maximized when $r_{\text{guess}} = \frac{r(i)}{\gamma}$ for the range $[0, \frac{r(i)}{\gamma}]$. Thus simplifying the above expression with this value of r_{guess} and the value of γ, with probability $\geq 1 - 2\delta'$, $r(S) \leq \gamma r(i)/10$. Thus with probability $\geq 1 - 2\delta'$,

$$\frac{r(i)}{r(S)+r(i)} \geq \frac{1}{1+\gamma/10} \geq \frac{6}{\gamma}.$$

By the Chernoff bound

$$\Pr\left(\frac{n_1(i)}{n_3} \leq \frac{5}{\gamma}\right) \leq \Pr\left(\frac{n_1(i)}{n_3} \leq \frac{r(i)}{r(S)+r(i)} - \frac{1}{\gamma}\right) \leq e^{-2n_3/\gamma^2}.$$

The lemma follows from the bound on n_3 and by the union bound total error probability $\leq 4\delta'$. ■

Note that the conditions in Lemmas 18 and 19 hold with probability $\geq 1 - \frac{2\delta}{5}$ by Lemmas 17 and 16. Furthermore, since we use all the steps at most $\log\log k$ times, by the union bound, the conclusion in Lemma 10 fails with probability $\leq \frac{2\delta}{5} + \log\log k \cdot 8\delta' = \frac{2\delta}{5} \leq \delta$.

C.3 Proof of Theorem 11

For the ease of readability, we divide the proof into several sub-cases. We first show that if $p = q$, then the algorithm returns `same` with high probability. Recall that for notational simplicity we redefine $\delta' = \frac{\epsilon\delta}{32m(n_4+1)\log\log k}$.

Lemma 20. *If $p = q$,* Closeness test *outputs* `same` *with error probability $\leq \delta$.*

Proof. Note that the algorithm returns `diff` only if any of the Test-Equals return `diff`. We call Test-Equal at most $\frac{16}{\epsilon} \cdot \log\log k \cdot m \cdot (n_4+1)$ times. The probability that at any time it returns an error is $\leq \delta'$. Thus by the union bound total error probability is $\leq \delta' 16 \log\log k \cdot m \cdot (n_4+1)/\epsilon \leq \delta$. ■

We now prove the result when $||p-q||_1 \geq \epsilon$. We first state a lemma showing that Prune-set ensures that set S does not have any elements $\geq 4r(i)$. The proof is similar to that of Lemmas 17 and 16 and hence omitted.

Lemma 21. *If i is α-heavy and β-approximable, then at any call of step 2 of* Assisted-closeness-test*, with probability $\geq 1 - \frac{2\delta}{5}$, if $r_{guess} \leq \frac{\gamma}{\beta}\frac{r(i)}{r(G_i)}$, then* Prune-test *never removes an element with probability $\leq 2r(i)$ and removes all elements with probability $\geq 4r(i)$.*

The proof when $||p-q||_1 \geq \epsilon$ is divided into two parts based on the probability of certain events. Let $\beta' = \frac{p(i)-q(i)}{p(i)+q(i)}$, $\beta'' = \frac{\alpha\beta}{128\gamma\log\frac{128\gamma}{\beta^2}}$. Let D denote the event such that an element j from G_i^c with $\left|\frac{p(j)-q(j)}{p(j)+q(j)} - \beta'\right| \geq \beta''$ and $r(j) \leq 4r(i)$ gets included in S. We divide the proof in two cases when $\Pr(D) \geq \frac{\alpha\beta^2}{128\gamma}$ and $\Pr(D) < \frac{\alpha\beta^2}{128\gamma}$.

Lemma 22. *Suppose $||p-q||_1 \geq \epsilon$. If i is α-heavy and β-approximable, $\frac{r(i)}{\gamma} \leq r_{guess} \leq \frac{\gamma}{\beta}\frac{r(i)}{r(G_i)}$, the conclusions in Lemma 21 hold , and $\Pr(D) \geq \frac{\alpha\beta^2}{128\gamma}$, then step $3(a)$ of* Assisted-closeness-test *returns* `diff` *with probability $\geq 1/5$.*

Proof. we then show that the following four events happen with high probability for at least one set $S \in \{S_1, S_2 \ldots S_m\}$.

- S includes a j such that $\left|\frac{p(j)-q(j)}{p(j)+q(j)} - \beta'\right| \geq \beta''$,$r(j) \leq 4r(i)$, $j \notin G_i$.
- $r(S) \leq r_{\text{guess}} + 8\sqrt{r(i)r_{\text{guess}}}$.
- j appears when S is sampled n_4 times.
- Test-equal returns `diff`.

Clearly, if the above four events happen then the algorithm outputs `diff`. Thus to bound the error probability, we bound the error probability of each of the above four events and use union bound. Probability that at least one of the sets contain an element j such that $\left|\frac{p(j)-q(j)}{p(j)+q(j)} - \beta'\right| \geq \beta''$,$r(j) \leq 4r(i)$, $j \notin G_i$ is

$$1 - (1 - \Pr(D))^m \geq 1 - e^{-\Pr(D)m} \geq \frac{5}{6}.$$

Let $S' = \{j \in S : r(j) \leq 4r(i)\}$. Observe that before pruning $\mathbb{E}[r(S')] \leq r_{\text{guess}}$ and $\text{Var}(r(S')) \leq 4r(i)r_{\text{guess}}$. Hence by the Chebyshev bound with probability $\geq 1 - 1/16$,

$$r(S') \leq r_{\text{guess}} + 8\sqrt{r(i)r_{\text{guess}}},$$

After pruning, $r(S)$ contains only elements of probabilities from S'. Hence with probability $\geq 1 - 1/16$, $r(S) \leq r_{\text{guess}} + 8\sqrt{r(i)r_{\text{guess}}}$. Probability that this element does appear when sampled n_4 times is

$$1 - \left(1 - \frac{r(j)}{r(S)}\right)^{n_4} \geq 1 - \left(1 - \frac{r(i)}{r_{\text{guess}}(1 + 8\sqrt{r(i)/r_{\text{guess}}})}\right)^{n_4} \geq 1 - \left(1 - \frac{\alpha\beta}{9\gamma}\right)^{n_4} \geq \frac{5}{6}.$$

Since $\left|\frac{p(j)-q(j)}{p(j)+q(j)} - \beta'\right| \geq \beta''$,$r(i) \leq r(j) \leq 4r(i)$ by Lemma 14 the chi-squared distance is

$$\geq (\beta'')^2 \frac{r(j)r(i)}{4(r(i)+r(j))^2} \geq \frac{(\beta'')^2}{25}.$$

Thus by Lemma 1, algorithm outputs `diff` with probability $1 - \delta'$. By the union bound the total error probability $\leq 1/6 + 1/16 + +1/6 + \delta' \leq 4/5$. ■

Lemma 23. *Suppose $||p - q||_1 \geq \epsilon$. If i is α-heavy and β-approximable, $\frac{r(i)}{\gamma} \leq r_{guess} \leq \frac{\gamma}{\beta}\frac{r(i)}{r(G_i)}$, the conclusions in Lemma 21 hold , and $\Pr(D) < \frac{\alpha\beta^2}{128\gamma}$, then step 3(b) of* Assisted-closeness-test *returns `diff` with probability $\geq 1/5$.*

We show that the following four events happen with high probability for at least some set $S \in \{S_1, S_2, \ldots, S_m\}$. Let $S' = S \cap G_i$ and $G_i' = G_i \setminus \{i\}$. Let

$$Z = r(S')\left|\beta' - \frac{p(S') - q(S')}{p(S') + q(S')}\right| = \frac{|\beta'(p(S') + q(S')) - (p(S') - q(S'))|}{2}.$$

- $Z \geq r_{\text{guess}}|(\beta'(p(G_i') + q(G_i')) - p(G_i') + q(G_i'))|/4$.
- $r(S) \leq 8(r(i) + r_{\text{guess}}) \log \frac{128\gamma}{\alpha\beta^2}$.
- Event D does not happen.
- Test-equal outputs `diff`.

Clearly if all of the above events happen, then the test outputs `diff`. We now bound the error probability of each of the events and use union bound. Since none of the elements in S' undergo pruning, the value of Z remains unchanged before and after pruning. Thus any concentration

inequality for Z remains the same after pruning. We now compute the expectation and variance of Z and use Paley Zigmund inequality.

$$\begin{aligned}\mathbb{E}[Z] &= \mathbb{E}[|\beta'(p(S') + q(S')) - p(S') + q(S')|]/2 \\ &\geq |\mathbb{E}(\beta'(p(S') + q(S')) - p(S') + q(S'))|/2 \\ &= r_{\text{guess}}|(\beta'(p(G'_1) + q(G'_i)) - p(G'_i) + q(G'_i))|/2,\end{aligned}$$

where the inequality follows from convexity of $|\cdot|$ function. Let $\mathbb{1}(j, S')$ denote the event that $j \in S'$. The variance is lower bounded as

$$\begin{aligned}\text{Var}(Z) &= \mathbb{E}[Z^2] - (\mathbb{E}[Z])^2 \\ &= \mathbb{E}[(\beta'(p(S') + q(S')) - p(S') + q(S'))^2]/4 - \mathbb{E}^2[|\beta'(p(S') + q(S')) - p(S') + q(S')|]/4 \\ &\overset{(a)}{\leq} \mathbb{E}[(\beta'(p(S') + q(S')) - p(S') + q(S'))^2]/4 - \mathbb{E}^2[\beta'(p(S') + q(S')) - p(S') + q(S')]/4 \\ &= \text{Var}(\beta'(p(S') + q(S')) - p(S') + q(S'))/4 \\ &\overset{(b)}{=} \sum_{j \in G'_i} \text{Var}(\mathbb{1}(j, S')(\beta'(p(j) + q(j)) - p(j) + q(j))^2/4 \\ &\leq \sum_{j \in G'_i} \mathbb{E}[\mathbb{1}(j, S')](\beta'(p(j) + q(j)) - p(j) + q(j))^2/4 \\ &= \sum_{j \in G'_i} r_{\text{guess}}(\beta'(p(j) + q(j)) - p(j) + q(j))^2/4 \\ &\leq \max_{j' \in G'_i} |\beta'(p(j') + q(j')) - p(j') + q(j')| \cdot r_{\text{guess}} \sum_{j \in G'_i} |\beta'(p(j) + q(j)) - p(j) + q(j)|/4 \\ &\overset{(c)}{\leq} 4r(i) \cdot r_{\text{guess}} r(G'_i).\end{aligned}$$

(a) follows from the bound on expectation. (b) follows from the independence of events $\mathbb{1}(j, S)$. (c) follows from the fact that $p(j) + q(j) = 2r(j) \leq 2r(i)$, $|\beta'| \leq 1$ and $\sum_i r(j) \leq r(G'_i)$. Hence by the Paley Zygmund inequality,

$$\begin{aligned}\Pr(Z \geq r_{\text{guess}}|(\beta'(p(G'_1) + q(G'_i)) - p(G'_i) + q(G'_i))|/4) &\geq \Pr(Z \geq \mathbb{E}[Z]/2) \\ &\geq \frac{1}{4}\frac{\mathbb{E}^2[Z]}{\text{Var}(Z) + \mathbb{E}^2[Z]} \\ &\geq \frac{1}{4}\frac{\mathbb{E}^2[Z]}{4r(G'_i)r(i)r_{\text{guess}} + \mathbb{E}^2[Z]}.\end{aligned}$$

Since i is β-approximable, by convexity

$$\frac{r(G'_i)}{r(G_i)}\left|\frac{p(i) - q(i)}{p(i) + q(i)} - \frac{p(G'_i) - q(G'_i)}{p(G'_i) + q(G'_i)}\right| \geq \left|\frac{p(i) - q(i)}{p(i) + q(i)} - \frac{p(G_i) - q(G_i)}{p(G_i) + q(G_i)}\right| \geq \beta.$$

Hence,

$$r(G'_i)\left|\frac{p(i) - q(i)}{p(i) + q(i)} - \frac{p(G'_i) - q(G'_i)}{p(G'_i) + q(G'_i)}\right| \geq r(G_i)\beta.$$

Thus $\mathbb{E}[Z] \geq r_{\text{guess}} r(G_i)\beta$ and

$$\begin{aligned}
\Pr(Z \geq r_{\text{guess}} | (\beta'(p(G'_1) + q(G'_i)) - p(G'_i) + q(G'_i))|/4) &\geq \frac{1}{4} \frac{(r_{\text{guess}} r(G_i)\beta)^2}{4r(i) r_{\text{guess}} r(G'_i) + (r_{\text{guess}} r(G_i)\beta)^2} \\
&\geq \frac{1}{4} \frac{(r_{\text{guess}} r(G_i)\beta)^2}{2\max(4r(i) r_{\text{guess}} r(G'_i), (r_{\text{guess}} r(G_i)\beta)^2)} \\
&= \frac{1}{8} \min\left(1, \frac{(r_{\text{guess}} r(G_i)\beta)^2}{4r(i) r_{\text{guess}} r(G'_i)}\right) \\
&\geq \frac{r(G_i)\beta^2}{32\gamma} \\
&\geq \frac{\alpha\beta^2}{32\gamma}. \qquad (4)
\end{aligned}$$

By Lemma 15, with probability $\geq 1 - \frac{\alpha\beta^2}{128\gamma}$,

$$r(S) \leq r_{\text{guess}} + \sqrt{8 r_{\text{guess}} r(i) \log \frac{128\gamma}{\alpha\beta^2}} + 4r(i) \log \frac{128\gamma}{\alpha\beta^2} \leq 8(r(i) + r_{\text{guess}}) \log \frac{128\gamma}{\alpha\beta^2}. \qquad (5)$$

Let $S'' = S \setminus S'$. If event D has not happened then for all elements $j \in S''$, $\left|\beta' - \frac{p(j)-q(j)}{p(j)+q(j)}\right| \leq \beta''$ and hence

$$\left|\frac{p(i) - q(i)}{p(i) + q(i)} - \frac{p(S'') - q(S'')}{p(S'') + q(S'')}\right| \leq \beta''. \qquad (6)$$

Combining the above set of equations,

$$\begin{aligned}
\left|\frac{p(i) - q(i)}{p(i) + q(i)} - \frac{p(S) - q(S)}{p(S) + q(S)}\right| &\overset{(a)}{\geq} \frac{r(S')}{r(S)} \left|\frac{p(i) - q(i)}{p(i) + q(i)} - \frac{p(S') - q(S')}{p(S') + q(S')}\right| - \frac{r(S'')}{r(S)} \left|\frac{p(i) - q(i)}{p(i) + q(i)} - \frac{p(S'') - q(S'')}{p(S'') + q(S'')}\right| \\
&\overset{(b)}{\geq} \frac{Z}{r(S)} - \beta'' \\
&\overset{(c)}{\geq} \frac{2r(G'_i) r_{\text{guess}}}{4r(S)} \left(\left|\beta' - \frac{p(G'_i) - q(G'_i)}{p(G_i) + q(G_i)}\right|\right) - \beta'' \\
&\geq \frac{r(G_i) r_{\text{guess}} \beta}{2r(S)} - \beta'' \\
&\overset{(d)}{\geq} \frac{r(G_i) r_{\text{guess}} \beta}{8r(S)}.
\end{aligned}$$

(a) follows from convexity and the fact that $|a + b| \geq |a| - |b|$, (b) follows from Equation (6), and

(c) follows from Equation (4). (d) follows from the fact that

$$\begin{aligned}
\beta'' = \frac{\alpha\beta}{128\gamma \log \frac{128\gamma}{\alpha\beta^2}} &\leq \frac{\beta}{128 \log \frac{128\gamma}{\alpha\beta^2}} \min\left(\frac{\alpha}{\gamma}, \alpha\right) \\
&\leq \frac{\beta}{128 \log \frac{128\gamma}{\alpha\beta^2}} \min\left(\frac{r(G_i) r_{\text{guess}}}{r(i)}, r(G_i)\right) \\
&= \frac{r(G_i) r_{\text{guess}} \beta}{64 \log \frac{128\gamma}{\alpha\beta^2} \cdot 2\max(r(i), r_{\text{guess}})} \\
&\leq \frac{r(G_i) r_{\text{guess}} \beta}{64 \log \frac{128\gamma}{\alpha\beta^2} (r(i) + r_{\text{guess}})} \\
&\leq \frac{r(G_i) r_{\text{guess}} \beta}{8 r(S)}.
\end{aligned}$$

Thus by Lemma 14, chi-squared distance is lower bounded by

$$\begin{aligned}
\left(\frac{r(G_i) r_{\text{guess}} \beta}{8 r(S)}\right)^2 \frac{r(i) r(S)}{4(r(i) + r(S))^2} &= \frac{r_{\text{guess}}^2 \beta^2 r(i) r^2(G_i)}{2^8 r(S)(r(i) + r(S))^2} \\
&\overset{(a)}{\geq} \frac{r_{\text{guess}}^2 \beta^2 r(i) r^2(G_i)}{2^{20}(r(i) + r_{\text{guess}})^3 \log^3 \frac{128\gamma}{\alpha\beta^2}} \\
&\geq \frac{r_{\text{guess}}^2 \beta^2 r(i) r^2(G_i)}{2^{20} \cdot 8 \max(r^3(i), r_{\text{guess}}^3) \cdot \log^3 \frac{128\gamma}{\alpha\beta^2}} \\
&= \frac{r^2(G_i)\beta^2}{2^{23} \log^3 \frac{128\gamma}{\alpha\beta^2}} \min\left(\frac{r(i)}{r_{\text{guess}}}, \frac{r_{\text{guess}}^2}{r^2(i)}\right) \\
&\overset{(b)}{\geq} \frac{r^2(G_i)\beta^2}{2^{23} \log^3 \frac{128\gamma}{\alpha\beta^2}} \min\left(\frac{\beta r(G_i)}{\gamma}, \frac{r^2(G_i)}{\gamma^2 r^2(G_i)}\right) \\
&\geq \frac{\alpha^3\beta^3}{2^{23}\gamma^2 \log^3 \frac{128\gamma}{\alpha\beta^2}}.
\end{aligned}$$

(a) follows from Equation (5) and (b) follows from bounds on r_{guess}. Thus with probability $\geq 1 - \frac{\alpha\beta^2}{128\gamma}$, TEST-EQUAL outputs `diff`. By the union bound, the error probability for an $S \in \{S_1, S_2, \ldots S_m\}$ is $\leq 1 - \frac{\alpha\beta^2}{32\gamma} + \frac{\alpha\beta^2}{128\gamma} + \frac{\alpha\beta^2}{128\gamma} + \delta' \leq 1 - \frac{\alpha\beta^2}{128\gamma}$. Since we are repeating it for m sets, the probability that it outputs `diff` is

$$\geq 1 - \left(1 - \frac{\alpha\beta^2}{128\gamma}\right)^m \geq 1 - e^{\frac{\alpha\beta^2 m}{128\gamma}} \geq \frac{1}{5}.$$

Theorem 11 follows from Lemma 20 for the case $p = q$. If $||p - q||_1 \geq \epsilon$, then it follows from 9 (finds a good tuple), 10 (finds good approximation of r_{guess}), 21 (pruning), and 22 ($\Pr(D)$ is large), and 23 ($\Pr(D)$ is small). By Lemma 20 success probability when $p = q$ is $\geq 1 - \delta$. The success probability when $||p - q||_1 \geq \epsilon$ is at least the probability that we pick a good-tuple (i, β, α) and the success probability once a good tuple is picked (sum of errors in Lemmas 10, 21 + maximum

of errors in Lemmas 22 and 23) which can be shown to be $1/5 \cdot (1/5 - \delta - 2\delta/5) \geq 1/30$. We now analyze the number of samples our algorithm uses.

We first calculate the number of samples used by Assisted-closeness-test. Step 2 calls Prune-set m times and each time Prune-set uses $n_1 n_2$ samples. Hence, step 2 uses mn_1n_2 samples. Step 3(a) uses $mn_4 \cdot \widetilde{\mathcal{O}}(\beta''^{-2})$ and step 3(b) uses $m \cdot \widetilde{\mathcal{O}}(\epsilon^{-3})$. Hence, the total number of samples used by Assisted-closeness-test is

$$mn_1n_2 + mn_4 \cdot \widetilde{\mathcal{O}}(\beta''^{-2}) + m \cdot \widetilde{\mathcal{O}}(\epsilon^{-3}) = \widetilde{\mathcal{O}}\left(\alpha^{-1}\beta^{-2}\epsilon^{-1} + \alpha^{-1}\beta^{-2}\epsilon^{-1}\epsilon^{-2} + \alpha^{-1}\beta^{-2}\epsilon^{-3}\right) = \widetilde{\mathcal{O}}\left(\beta^{-1}\epsilon^{-4}\right).$$

Thus each Assisted-closeness-test uses $\widetilde{\mathcal{O}}(\epsilon^{-4}\beta^{-1})$ samples. Hence, the number of samples used by Binary-search is

$$\leq \widetilde{\mathcal{O}}\left(\log\log k \left(n_1n_2 + \epsilon^{-4}\beta^{-1} + n_3\right)\right) = \widetilde{\mathcal{O}}\left(\frac{\log\log k}{\epsilon^4\beta}\right).$$

Since Closeness-test calls Binary-search for $16/\epsilon$ different tuples. Hence, the sample complexity of closeness test is

$$\frac{16}{\epsilon} + \sum_{j=1}^{16/\epsilon} \widetilde{\mathcal{O}}\left(\frac{\log\log k}{\epsilon^4\beta_j}\right) = \widetilde{\mathcal{O}}\left(\frac{\log\log k}{\epsilon^5}\right).$$